Translation of quantitative MRI analysis tools for clinical neuroradiology application

Olivia Goodkin MBBS

Thesis submitted to University College London for the degree of Doctor of Philosophy

August 2021

I, Olivia Goodkin, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Dr Sjoerd Vos assisted with image processing for chapter 3. Dr Ferran Prados and Hugh Pemberton assisted with image processing for chapter 4. Dr Carole Sudre and Dr Ferran Prados assisted with image processing for chapter 5. Dr Ravi Das assisted with statistical analysis for chapters 3 and 4.

Olivia Goodkin

Abstract

Quantification of imaging features can assist radiologists by reducing subjectivity, aiding detection of subtle pathology, and increasing reporting consistency. Translation of quantitative image analysis techniques to clinical use is currently uncommon and challenging. This thesis explores translation of quantitative imaging support tools for clinical neuroradiology use. I have proposed a translational framework for development of quantitative imaging tools, using dementia as an exemplar application. This framework emphasises the importance of clinical validation, which is not currently prioritised. Aspects of the framework were then applied to four disease areas: hippocampal sclerosis (HS) as a cause of epilepsy; dementia; multiple sclerosis (MS) and gliomas.

A clinical validation study for an HS quantitative report showed that when image interpreters used the report, they were more accurate and confident in their assessments, particularly for challenging bilateral cases. A similar clinical validation study for a dementia reporting tool found improved sensitivity for all image interpreters and increased assessment accuracy for consultant radiologists. These studies indicated benefits from quantitative reports that contextualise a patient's results with appropriate normative reference data. For MS, I addressed a technical translational challenge by applying lesion and brain quantification tools to standard clinical image acquisitions which do not include a conventional T1-weighted sequence. Results were consistent with those from conventional sequence inputs and therefore I pursued this concept to establish a clinically applicable normative reference dataset for development of a quantitative reporting tool for clinical use. I focused on current radiology reporting of gliomas to establish which features are commonly missed and may be important for clinical management decisions. This informs both the potential utility of a quantitative report for gliomas and its design and content.

I have identified numerous translational challenges for quantitative reporting and explored aspects of how to address these for several applications across clinical neuroradiology.

Impact statement

Image quantification reporting techniques are successfully developed and applied in the research setting but there are significant barriers to their translation for clinical use and to date they are not widely implemented. Here I present a translational framework to facilitate development of clinically focused quantification solutions. The framework is designed for widespread adoption by healthcare technology developers and is relevant to all stakeholders, including academic institutions, commercial companies, and clinical adopters.

A major unknown is the impact of these tools in the hands of clinical end-users. The multi-rater clinical validation studies that I have conducted and present in this thesis have not been performed before for clinical neuroradiology reporting tools. They have involved radiologists and image analysts from a variety of clinical institutions across the United Kingdom and Europe. These studies inform radiologists on the scope and impact they can expect from these tools and also set the standard for equivalent reporting tools to provide similar clinical validation evidence, for example those offered by commercial enterprises.

As a consequence of establishing clinical validity, the quantitative reporting tool for HS has already been integrated into clinical use in the local tertiary neuroradiology centre at the National Hospital for Neurology and Neurosurgery, Queen Square, London and the Chalfont Centre for Epilepsy, Chalfont St Peter. The translation of this tool has required education and engagement across the clinical department, including radiographers, clinical scientists, and radiologists alike. In-use evaluation of the reporting tool is ongoing, with benefits to radiologists, clinical referrers, and ultimately their patients being anticipated. Successful local adoption paves the way for further integration into more imaging departments in other institutions.

The MS work that I have undertaken has cultivated a strong and ongoing collaboration with MAGNIMS, the Magnetic Resonance Imaging in Multiple Sclerosis pan-European consortium, which pioneers MRI implementation for MS with international impact. The MS projects that I have presented in this thesis are supported by MAGNIMS and use their multi-centre data. The

projects are discussed at twice-yearly plenary meetings where I am able to receive feedback and expert guidance from leaders in the field and the results of my research can be disseminated.

Much of the work presented in this thesis is associated with publications in peer-reviewed journals. I have presented results of this research in oral presentations at national and international conferences including the European Society of Neuroradiology (ESNR) annual conference in 2019. The MS work was awarded an early career scholarship prize at the joint meeting of the American and European Committees for Treatment and Research in Multiple Sclerosis (ACTRIMS-ECTRIMS) in 2020.

I have ongoing support from the UCLH Biomedical Research Centre to continue to develop the work that I have conducted to date by undertaking postdoctoral research, reflecting the priority that is being placed on translational research in the field of healthcare technology.

Acknowledgements

I am grateful to the NIHR UCLH Biomedical Research Centre for supporting my research. I look forward to continuing my work within the Healthcare Engineering and Imaging theme.

I would like to thank all the research study subjects and patients whose imaging I have used to carry out this research, it simply would not have been possible without them.

Thank you to my supervisors, Professor Frederik Barkhof, Dr John Thornton, and Dr Sjoerd Vos. I am so grateful for their guidance and support throughout. Thank you to Frederik for giving me this opportunity, helping me to develop my confidence as a researcher and for always encouraging me to strive to reach for goals which may appear daunting at first. Thank you to John for his constant encouragement and sound advice and to Sjoerd for his close guidance and friendship, it was a pleasure to work together particularly on the HS project.

I am grateful to Professor Tarek Yousry for his mentorship and for always taking an active interest in me and my future development. His encouragement, warmth and detailed and incisive feedback have all been much appreciated.

Dr Ferran Prados has been like an informal supervisor to me, and I am so grateful for his unwavering presence, expertise, patience and friendship, without whose input much of this research would not have been possible. My particular thanks go to Hugh Pemberton; I am very lucky to have taken this journey with such an excellent colleague and to have made a true friend for life.

I would like to thank James Moggridge at the Lysholm Department of Neuroradiology for his fantastic dedication to achieving clinical translation of this work. I would also like to thank Dr Sotirios Bisdas, Professor Rolf Jager, Annie Papadaki, Lisa Strycharczuk and Soledad Delgado Dotor from the department for their important input.

Thanks go to everyone at the Queen Square MS Centre, in particular to Professor Olga Ciccarelli for her invaluable guidance and mentorship,

Professor Ahmed Toosy for giving me my first experience in research as a BSc student back in 2009, Dr Arman Eshaghi, Dr Carmen Tur, Dr Lukas Haider, Dr Sara Collorone, Marios Yiannakas, Jon Steel, Charlotte Burt and Tina Holmes.

I am grateful to colleagues at CMIC, especially Dr James Cole and Dr Carole Sudre who have been significant guiding figures in helping me to shape my research and to develop as a researcher. My sincere thanks also to Professor Danny Alexander, Professor Gary Zhang and Dr Baris Kanber.

It has been fantastic to collaborate recently with some excellent early career researchers and peers: Jiaming Wu, Dr Jordan Coleman, and Dr Giuseppe Pontillo, and I am looking forward to continuing to work with them.

I would also like to thank collaborators from elsewhere at UCL and further afield, including Dr Jo Barnes, Professor Nick Fox, Dr Gavin Winston, Professor John Duncan, Professor Meike Vernooij, and Dr Jorge Cardoso who supervised me at the beginning of my PhD. I am immensely grateful to the MAGNIMS steering committee and all its contributors.

Finally, I would like to thank my wonderful family and friends, they know who they are, and I could not have done this without them. The past 18 months have been particularly challenging as we have navigated the pandemic, during which time I returned to assist on the wards at UCLH through the first peak. They gave me the strength I have needed in such strange times as they have at many other times too.

Publications associated with this thesis

Goodkin O, Pemberton HG, Vos SB, Prados F, Sudre CH, Moggridge J, Cardoso MJ, Ourselin S, Bisdas S, White M, Yousry T, Thornton J, Barkhof F, 2019. The quantitative neuroradiology initiative framework: application to dementia. Br J Radiol. 92(1101). doi:10.1259/bjr.20190365.

Goodkin O, Pemberton HG, Vos SB, Prados F, Das RK, Moggridge J, De Blasi B, Bartlett P, Williams E, Campion T, Haider L, Pearce K, Bargalló N, Sanchez E, Bisdas S, White M, Ourselin S, Winston GP, Duncan JS, Cardoso J, Thornton JS, Yousry TA, Barkhof F, 2021. Clinical evaluation of automated quantitative MRI reports for assessment of hippocampal sclerosis. Eur Radiol. 31(1):34-44. doi:10.1007/s00330-020-07075-2.

Pemberton HG*, Goodkin O*, Prados F, Das RK, Vos SB, Moggridge J, Coath W, Gordon E, Barrett R, Schmitt A, Whiteley-Jones H, Burd C, Wattjes MP, Haller S, Vernooij MW, Harper L, Fox NC, Paterson RW, Schott JM, Bisdas S, White M, Ourselin S, Thornton JS, Yousry TA, Cardoso MJ, Barkhof F, 2021. Automated quantitative MRI volumetry reports support diagnostic interpretation in dementia: a multi-rater, clinical accuracy study. Eur Radiol. 31(7):5312-5323. doi:10.1007/s00330-020-07455-8. *Joint first authors.

Goodkin O, Prados F, Vos SB, Pemberton H, Collorone S, Hagens MHJ, Cardoso MJ, Yousry TA, Thornton JS, Sudre CH, Barkhof F, MAGNIMS study group, 2021. FLAIR-only joint volumetric analysis of brain lesions and atrophy in clinically isolated syndrome (CIS) suggestive of multiple sclerosis. Neuroimage Clin. 29:102542. doi:10.1016/j.nicl.2020.102542.

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1 Application of quantitative MRI biomarkers to clinical neuroradiology

1.1.1 Current clinical radiology practice

Clinical radiology practice currently relies upon the individual visual assessment and interpretation of an image or series of images by a clinically trained medical practitioner, usually a radiologist. The individual radiologist will provide a description of qualitative features of the image, and their overall impression of the presence and extent of pathology. The features that they describe are often difficult to define objectively, and the ability to differentiate appearances along a spectrum of normality and abnormality are heavily reliant on the radiologist's accumulated experience and knowledge of the bounds of normality, including anatomical and physiological variations. Their ability to judge the findings of a particular imaging examination is dependent upon their ability to reference it against previous cases they have seen and the training they have received. This allows for potential variability of interpretation between practitioners.

Semi-quantitative assessment has been adopted in several key areas in response to this. These include visual rating scales (VRS), defined as discrete categories which enable clearer communication of key findings, as well as the use of structured reporting systems which aim to standardise the content of radiology reports. These have been particularly successful in cancer reporting where reporting and data systems (RADS) are currently used to communicate important findings and facilitate clinical decisions, for example in prostate and breast cancer (Barentsz et al. 2016; Mercado 2014).

Semi-quantitative assessment provides an element of standardisation, however it remains a somewhat coarse classifier. Extraction of objectively quantifiable features from clinical images and providing a reference normative comparison against which to contextualise an individual patient's imaging has the possibility to transform the use of medical images, so that they can be interpreted not only as pictures but also analysed for their rich data content.

The standard visual assessment paradigm that accounts for the mainstay of current clinical imaging and possible opportunities for quantitative assessment is depicted in Figure 1-1.

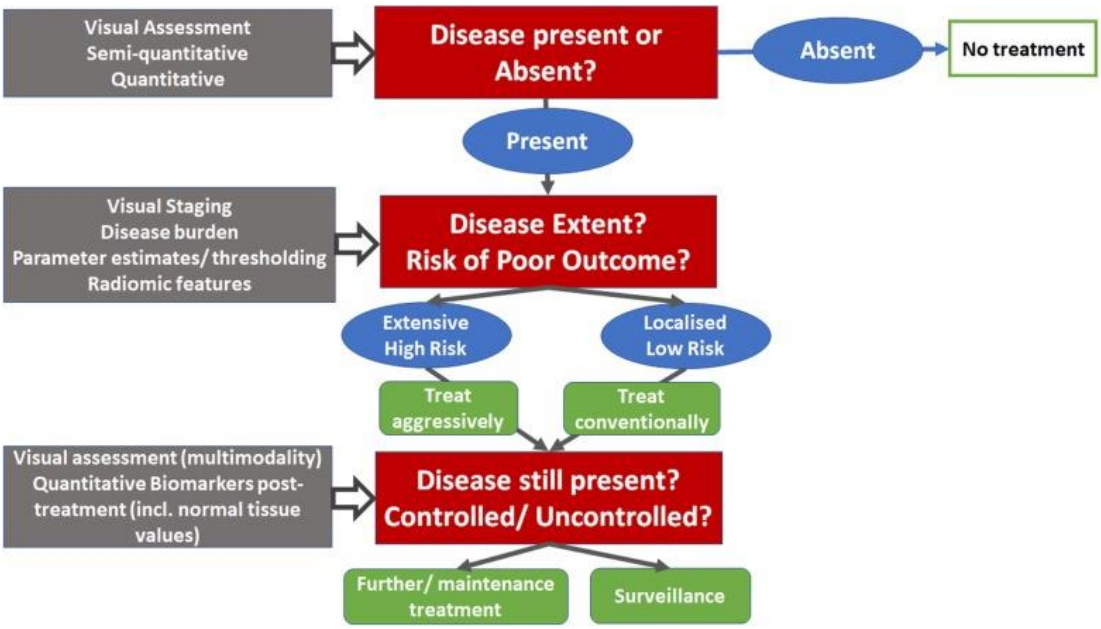


Figure 1-1. A schematic of how a quantitative imaging tool may be used to assess three crucial questions: diagnosis, disease extent, and disease progression. These are mostly evaluated by subjective visual assessment in current clinical radiological practice. Figure from deSouza et al. 2019.

1.1.2 Quantitative Imaging Biomarkers (QIBs)

Biomarkers are measurable objective indicators of a pathological or physiological state. They are characteristics that should be possible to measure accurately and precisely to provide a true representation of an aspect of a particular condition.

Quantitative imaging biomarkers, QIBs, are defined as objective characteristics derived from *in vivo* images, that can be measured, and can indicate either normal physiology, pathology, or response to a therapeutic intervention (Sullivan et al. 2015). They offer the potential for detection of key features relating to a pathological process with sensitivity and reproducibility not attainable by qualitative observations.

There are several requirements which a QIB should fulfil (Smits 2021). Their precision, accuracy and ‘trueness’ should be demonstrated. Precision refers to measurement variability, which consists of repeatability and reproducibility. Repeatability is the demonstration that consistent measurements are produced during repeated trials, whereas reproducibility is defined as demonstrating consistent measurements where there has been a change in operator, system or measurement device (Kessler et al. 2015). In the case of

imaging biomarkers, there are many different sources of potential variability, for example between populations and use of different image acquisition parameters. Accuracy refers to the performance of the QIB in a clinical setting, measured by sensitivity and specificity for the intended condition. Trueness is defined as the closeness of a QIB measurement to a certain true reference. This is often difficult to definitively establish, for example in cases where the reference is histological and therefore often not possible to determine *in vivo*. In other cases, an established reference may not exist.

Inherent to QIBs, unlike biological biomarkers such as those derived from blood analysis, is the fact that their extraction is heterogeneous. This is because an imaging feature is being interpreted as a surrogate for a real biological process, and this feature is being measured using complex and variable equipment, depending on factors like scanner design and the selected acquisition protocol.

The Radiological Society of North America (RSNA) has established the Quantitative Imaging Biomarker Alliance (QIBA) to promote the validation and translation of QIBs into clinical use. They highlight that quantitative imaging requires related elements that facilitate accurate and communicable use of QIBs to be addressed; that is the standardisation and optimisation of acquisition protocols, data analysis, how data is displayed, and the implementation of structured reporting (RSNA 2007).

Currently QIBs are yet to have a significant impact on clinical imaging routines, which is partially due to the paucity of widespread standardisation, as well as the lack of technical and clinical validation. The European Society of Radiology (ESR) have published guidelines to promote a standard approach to the validation of image acquisition and analysis methods (ESR 2020). They base QIB validation steps on the three central principles of precision, accuracy, and clinical relationship (Figure 1-2).

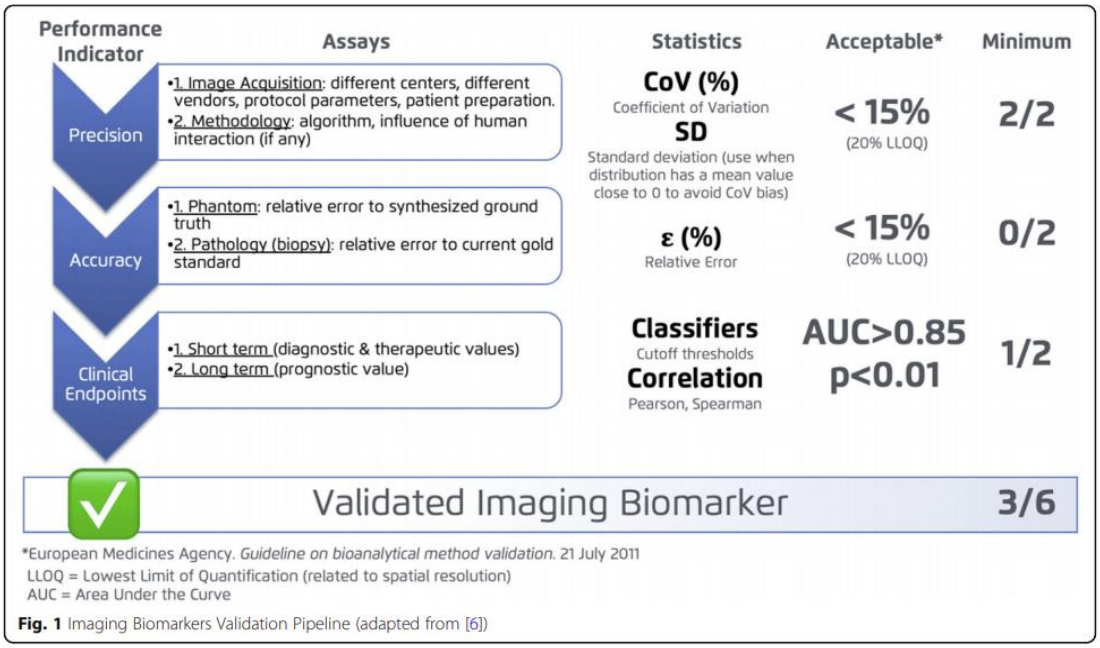


Figure 1-2. Methodological validation of a QIB. From the ESR Statement on the Validation of Imaging Biomarkers (ESR 2020).

They suggest that measurement precision should be tested by varying image acquisition conditions, for example testing in different centres, using different MRI field strengths or multiple scanner vendors, and applying various acquisition protocols. Analysis algorithms and software should likewise be tested in the hands of different operators and with any processing variability.

Accuracy may be established against appropriate synthetic phantoms or tissue samples; however the guidelines acknowledge that this may not always be possible. In these instances, clinical sensitivity and specificity may replace accuracy, as they show how strongly the QIB is related to a particular pathological or physiological process.

The third key principle, clinical relationship, refers in this technical validation context to the ability to show that the QIB is related to a disease status. This relationship may be to detection and diagnosis, or to the assessment of response to treatment, or to disease progression over time.

1.1.3 Clinical translation of QIBs

Treating technical validation of QIBs as equivalent to biological biomarkers sets a standard that should ensure that QIBs that reach clinical use are

thoroughly tested with the potential to make a positive clinical impact. However the validation requirements are not trivial to achieve, given the extent of variability of image acquisition and processing across the field (Smits 2021). Reproducibility of results is difficult to demonstrate in the clinical setting, where patients do not routinely have multiple scans at the same time-point to facilitate this analysis (deSouza et al. 2019).

Usually QIBs are first identified, extracted, and tested in group-level research studies, which may present a further translational challenge for the application to an individual patient. Research software applications are not designed with clinical application in mind, which presents several translational challenges. They are designed to process imaging data that has been acquired to specific research or clinical trial standards which often surpass clinical acquisitions both in terms of image quality and standardisation. They also do not engage with quality management and medical device regulatory mechanisms or clinical validation that should precede implementation in routine radiology practice. They are not designed to be easily embedded into hospital information technology (IT) infrastructure, for example for deployment within the picture archiving and communication system (PACS).

Considering all these significant implementation barriers, it is important to deconstruct the process of QIB extraction and analysis with the clinical environment in mind as the target destination, so that a translational framework for their use can be explored and implemented.

Several international efforts are being made to engage in earnest with the translational issues facing clinical QIB implementation. The previously mentioned RSNA's QIBA collaborates with the ESR's European Imaging Biomarkers Alliance (EIBALL) to provide guidelines for imaging standardisation in terms of acquisition and processing, as well as setting out validation procedures for implementation (ESR 2021b).

These efforts come in place of established quality assurance and control (QA and QC) procedures set by regulatory bodies, which do not yet exist for QIBs. However, QIBA and EIBALL themselves differ on the guidance thresholds for validation that they provide. Both agree on the importance of clearly stating the

context for any QIB quantification or validation, for example the population it was tested in, and the imaging parameters employed, to ensure that results are interpretable in their correct context.

For widespread adoption of a QIB to occur, comparability of results across different platforms and institutions must be demonstrated. Finally, once a QIB has been technically and clinically validated in multi-centre studies, its added value should be proven. Installation of new software and other necessary equipment including QC mechanisms, as well as training or provision of qualified staff to operate new processing pipelines come with significant costs. Cost effectiveness is an increasingly important concern in financially restricted healthcare systems. Cost-benefit analysis should consider the wider context of QIB use, including whether they are being used alongside biological biomarkers or are able to replace other tests. Prospective analysis should also consider whether the cost of QIB use was later offset by improved targeting of expensive therapeutic interventions.

Cost effectiveness may also be partly evidenced by increased efficiency and time benefit to radiologists who are facing ever-increasing workloads (Chen and Lexa 2017). This potential time saving should be weighed up against any additional time required for scanning or post-processing techniques for QIB analysis which may delay clinical interpretation. Additional training requirements for radiologists in QIB interpretation should also be considered (Smith 2011).

Technical and clinical validity of a QIB are not the only translational considerations to overcome. Clinical environments can be somewhat esoteric in their organisational procedures and structures, and attitudes towards adoption of new technologies may vary and compete with other financial investments (Strohm et al. 2020).

The NASSS framework (Non-adoption, Abandonment, Scale-up, Spread, and Sustainability) can be used to pinpoint the key factors affecting implementation of complex healthcare technologies (Greenhalgh et al. 2017). It considers seven domains: condition, technology, value proposition, adopter system, organisation, and wider institutional and social context. When this framework

was applied to Dutch healthcare use of a bone density measurement software solution (Strohm et al. 2020), researchers found that while users expected the application to be of significant added value, in terms of avoiding errors and automation of time-consuming manual tasks, there was scepticism regarding the technical consistency and reliability of software. Users expect applications to be fully integrated with their existing workflows and information technology (IT) systems, minimising additional steps required for the radiologist to engage with its results.

Attitudes towards QIB adoption were shown to be influenced by organisational openness to innovation. Promotion of innovation by clinical and management leadership is complemented by the presence of local departmental champions, who take a lead in educating and stimulating interest among their colleagues.

Some technical understanding amongst radiologists of the technology they are using has been shown to be important, not only for understanding the limitations of an application but also for creating trust in how results are reached (Rubin 2019). It is important that the output of a quantitative tool is transparent to the end-user. The absence of guidelines for best practice means that evaluation of an application in-use is often unstructured and lacking defined outcomes for assessment. It is also difficult to demonstrate to radiologists that the application is having a proven clinical impact beyond their anecdotal experience. Ultimately some radiologists may fear a perceived change in their role or anticipate their professional expertise or autonomy being undermined when asked to accommodate QIB technology into their workflow (Huisman et al. 2021). Likewise perceived acceptability of the technique to referring clinicians is also important as they may otherwise be hesitant to incorporate the computer-assisted results into their clinical decision making.

Clinical QIB translation requires development within a quality management framework and regulatory approval from a recognised body, for example the Conformité Européenne (CE). Regulatory approval currently depends more on demonstration of completing technical development stages than of benefit to

clinical practice, which may contribute to additional clinical adoption uncertainty.

1.2 Existing commercial quantitative reporting tools for clinical use

Several commercial tools exist for MRI QIB quantification in the routine clinical neuroradiology setting, predominantly in the form of quantitative reports for use in suspected dementia. These reports have received regulatory certification from CE and/or the Food and Drug Administration (FDA).

Despite their regulatory approval, the technical and clinical validation of these tools is variable, and it is often difficult for clinicians and clinical institutions to establish the evidence that supports their use. In particular, there is a paucity of evidence regarding their clinical application and impact on detection accuracy and clinical decision making in the hands of end-users.

The most established and widely used of these commercial tools, NeuroQuant, was developed by cortechs.ai (www.cortechslabs.com, cortechs.ai 2021), which received FDA clearance in 2006. Reporting tools commonly use automated brain segmentation algorithms and present single-subject results contextualised by normative reference data in a graphical report format. Most tools use atlas-based segmentation approaches while a minority claim to use deep learning methods, for example Mediaire (www.mediaire.de, mediaire 2021).

Methodologies are often only described as proprietary with no further details provided, making them difficult to independently assess. Several companies use modified versions of previously technically validated research methods, such as Freesurfer (Fischl 2012), used by cortechs.ai and ADM diagnostics (www.admdx.com, ADMdx 2021), Voxel Based Morphometry (VBM) (Good et al. 2001) used by jung diagnostics (www.jung-diagnostics.de, jung diagnostics 2021) and Geodesic Information Flows (GIF) (Cardoso et al. 2015), used by Brainminer (www.brainminer.co.uk, Brainminer 2021). They all provide volumetric quantification of the hippocampi and brain lobes, while some provide additional sub-regional volume measurements and ventricular volume.

It is also variable whether commercial tools offer longitudinal analyses in addition to cross-sectional reports. Those that do provide a longitudinal report either use an indirect measurement approach by calculating the difference between two cross-sectional reports, for example Quantib (www.quantib.com,

Quantib 2021), or a direct measurement approach such as the boundary shift integral (Freeborough and Fox 1997; Prados et al. 2015) or SIENA (Smith et al. 2002), including Cortechs.ai, Icometrix (www.icometrix.com, icometrix 2021) and Combinostics (www.cneuro.com, Combinostics 2021). All tools offer integration with clinical workflow platforms, i.e. the Picture Archiving and Communication System (PACS).

Some fundamental aspects of the reports are unfortunately opaque to the intended end-user. These include the composition of the reported normative reference data, and the details of the QC process. Normative data should ideally cover a representative range of age, ethnicity and gender to make it broadly applicable. If the reporting tool is intended for implementation across many different centres and to handle MRI data from a range of scanner vendors and field strengths, this heterogeneity should be replicated within the makeup of the normative reference data. Commercial tools that make this information difficult to access may limit their own implementation and applicability. In the same vein, it is important for the end-user to understand what QC is performed to ensure that scans are of adequate quality to enable the quantification tool to process it and produce a meaningful and reliable result.

Technical and clinical validation studies using these commercial volumetric reporting tools are variable in their quantity. Technical validation is commonly performed by comparing the commercial tool to results of manual segmentation or a benchmark research tool such as Freesurfer (Fischl, Sereno, and Dale 1999). More established companies such as CorTechs.ai and Icometrix have undergone the most technical validation in the literature. Cortechs.ai segmentation results have been compared against manual segmentations (Brewer et al. 2009; Brinkmann et al. 2019), as well as results from Freesurfer (Ochs et al. 2015; Yim et al. 2021) and FSL-FIRST in non-dementia conditions (Lyden et al. 2016; Pareto et al. 2019). Icometrix's dementia tool output has been compared to Freesurfer (Struyfs et al. 2020).

Clinical validation studies have only been performed for a minority of commercially available reporting tools. Hippocampal volumetric results from

NeuroQuant have been compared to visual rating scale classification for medial temporal lobe atrophy, i.e. the Scheltens scale (Scheltens et al. 1992), and they showed good correlation with the visual method (Min et al. 2017; Persson et al. 2018). NeuroQuant hippocampal measurements were also shown to discriminate AD from a non-dementia group of patients with subjective memory complaints or mild cognitive impairment (Persson et al. 2017).

Neuroreader, the product by Brainreader (www.brainreader.net, brainreader 2021) and an open-source research tool volBrain (Manjón and Coupé 2016) were compared to both automatic classification results by machine learning and to the visual classification of two radiologists, using patient data from a single memory clinic. Neuroreader displayed moderate accuracy compared to the gold standard and was inferior to neuroradiologists and machine learning classification (Morin et al. 2020). A direct comparison of the potential prognostic efficacy of Neuroreader and NeuroQuant showed they produced comparable results in a longitudinal standardised dataset (Tanpitukpongse et al. 2017).

A clinical accuracy study which used jung diagnostics' Biometrica reports compared two radiologists' diagnoses with and without the assistance of the quantitative reports in a clinical dataset (Hedderich et al. 2020). The presence of a quantitative report was shown to significantly improve differentiation of dementia patients from healthy controls for one of the raters who was the least experienced. It also improved both radiologists' ability to differentiate between dementia subtypes and improved inter-rater agreement. This is the only clinical validation study for a commercial tool to date that attempts to simulate the clinical workflow and assess the impact of integrating the quantitative report with the radiologist's visual inspection.

While there is increasing interest in QIB use and their potential clinical utilisation, there is a general noticeable lack of clinical validation for quantitative neuroradiology reporting tools. Their impact when being applied by intended end-users, in studies which use heterogeneous clinical-grade

imaging data representing the patient population, should be prioritised to facilitate the translation of QIB reporting tools to clinical use.

1.3 Aims of this thesis

In this thesis I will explore the translational pathway required to successfully introduce quantitative image analysis techniques into clinical application using routine standard of care images. I will set out that a translational QIB pathway should aim to automatically derive robust and objective QIBs that are related to a disease state, using a validated technique at the level of the individual subject, and present results to the clinical practitioner in an accessible and interpretable format. I hypothesise that the provision of interpretable QIBs within the clinical radiology workflow could increase the clinician's assessment accuracy and confidence and reduce disagreement between independent image readers.

I aim to define a clinical translational framework for the development of neuroradiology QIB reporting tools, and then apply elements of the framework to four neuroradiological disease areas: hippocampal sclerosis (HS) in epilepsy, dementia, multiple sclerosis (MS), and gliomas. I will identify relevant QIBs that can be extracted from standard of care image acquisitions for these conditions and explore their technical and clinical validation and subsequent integration with the clinical workflow. In the cases of HS and dementia, I will focus on the clinical validation of quantitative reporting tools for these conditions by studying the impact of the reports on end-user accuracy and confidence when used as adjuncts to the standard visual assessment. For MS, I will apply quantitative methods to clinical MS image protocols that usually require non-standard of care imaging sequences, which limits their translation. I will apply the results to establishing a clinically applicable reference dataset for an MS reporting tool. For gliomas, I will concentrate on establishing the clinical need for an image quantification tool and work towards its design by studying the content and outcomes of current neuroradiology reporting.

By applying elements of the translational pathway to these different disease areas, I aim to investigate the areas where quantitative reporting tools may add value to clinical neuroradiology, quantify what these benefits may be, and highlight the challenges that may hamper their clinical translation.

2. A translational framework for quantitative MRI analysis

2.1 The Quantitative Neuroradiology Initiative (QNI) Framework

I will now introduce a translational framework to address the clinical translation of QIBs to the clinical neuroradiology setting.

The intention of the QNI framework is to provide a structure through which the automated image quantification can be efficiently adopted by and implemented within clinical neuroradiology practice. The focus is the design and delivery of clinically relevant, technically and clinically validated, user-friendly reports of regional and/or global brain characteristics. These reports present information at the individual patient level with the provision of contextual reference data, that are fully integrated into the clinical workflow.

The framework proposes the following six steps for adoption of specific quantitative neuroradiological tools into routine clinical practice, which will be described in further detail below and in a pictorial representation in Figure 2-1.

1. Establish the area of clinical need and the appropriate proven quantitative imaging biomarker(s).
2. Develop a method for automated analysis of the identified QIB(s), which includes algorithm development and compilation of suitable reference data.
3. Communicate the results of QIB analysis via an intuitive and accessible quantitative reporting tool.
4. Perform technical and clinical validation of the proposed reporting tool
5. Integrate the developed analysis pipeline into the clinical reporting workflow.
6. Perform in-use evaluation.

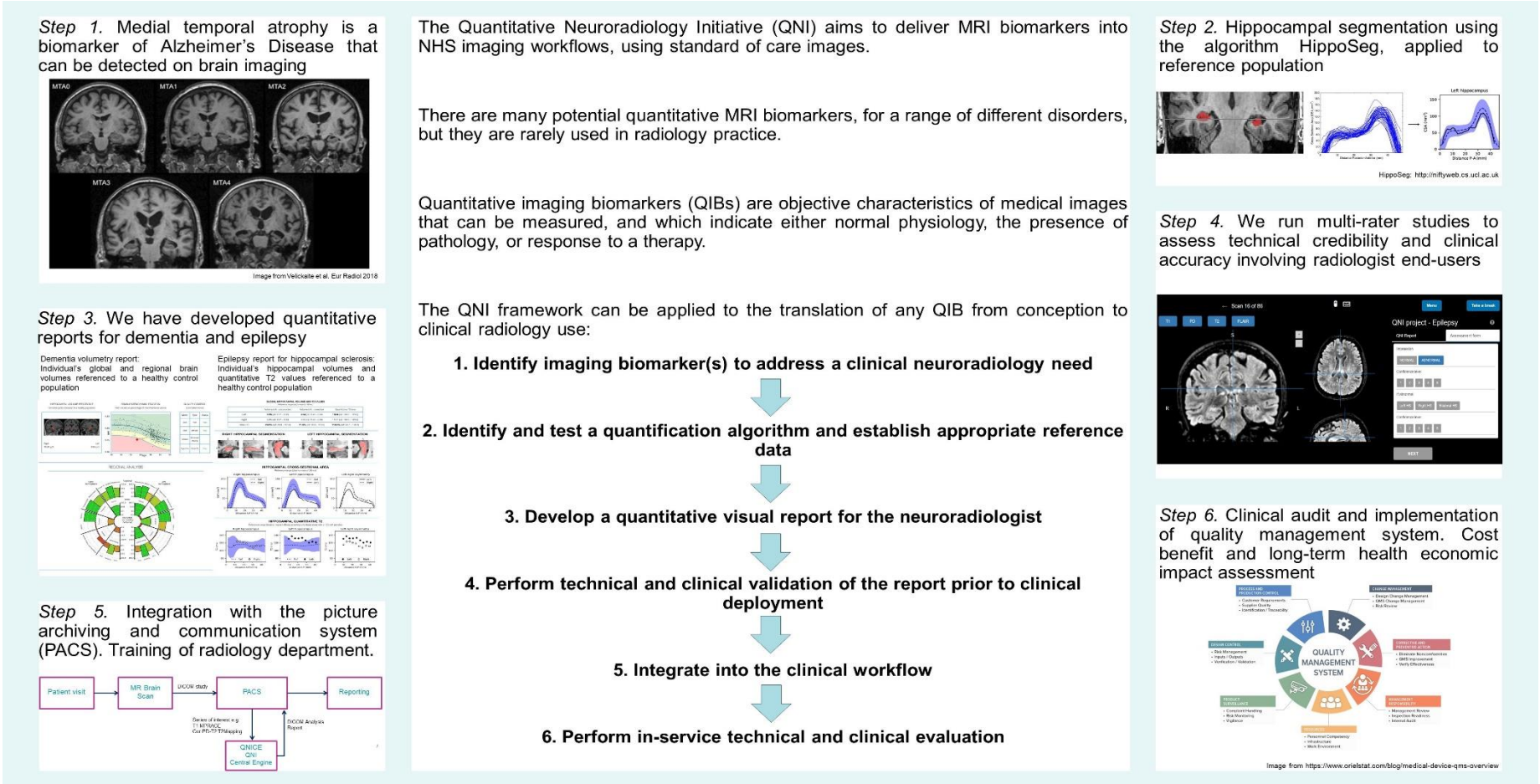


Figure 2-1. An overview of the QNI framework with pictorial examples given for each of the six translational steps. Steps 2, 3 and 4 use examples from studies that follow in this thesis.

2.1.1 Step 1. Establishing a clinical need and the appropriate QIB

The basis for accepting a particular imaging parameter as a correlate of the pathological process of interest should be demonstrated, and its implications for contribution to patient management should be considered. To this end, the chosen imaging biomarker should have a demonstrable physiological or pathological correlate, display sensitivity to the disease state, and be reproducible in its results. There should be a discernible additional benefit provided by the use of the QIB in addition to the accepted clinical routine. This benefit may be seen in increasing accuracy of diagnosis, prediction of disease course or treatment response, or surveillance for change over time. QIB relationship to a disease state as well as the particular benefit of measuring it should have been established in the research setting prior to its selection as a candidate for clinical translation. Evidence may include use of the QIB in a clinical trial, and validation against clinical and pathological data.

2.1.2 Step 2. Developing a method for automated analysis

An algorithm for automated quantitative analysis of the identified QIB(s) should be chosen or developed. This may be derived from those that have previously been used in the research setting and which may need to be adapted in some way to process clinical MRI data. The technique may have been applied to one disease area in the research setting and require alteration for application to a different disease area, for example by training an algorithm with new input data.

At this stage disease-specific normative reference data should be identified and processed in order to provide context for comparison of an individual patient's findings. Each disease area and biomarker of interest may require its own normative dataset to accurately contextualise the results. Ideally, a source of high volume, age-matched, generalisable normative data that adequately reflects the heterogeneity of clinical imaging data should be established to provide a range of normative values which can be taken to represent the normal range of values that would be encountered in the population. With large reference populations, generalisability of normative reference ranges should be less impaired by factors which introduce statistical noise, which include

scanner and imaging acquisition parameter differences, as well as patient demographic heterogeneity.

2.1.3 Step 3. Communicating QIB results

A report of the quantitative analysis should be produced. This should present data in a clinically and visually meaningful way, to ensure that the radiologist is provided with clear and accessible information that can be easily assimilated into their analysis, and their final report. Considered report design is important so that quantitative information is organised simply and intuitively. The availability of quantitative information may facilitate increased adoption of standardised or structured reporting that is already being implemented for qualitative analysis in some disease areas.

2.1.4 Step 4. Technical and clinical validation

The combined processes of automated algorithm analysis and report generation, which can be referred to as the processing 'pipeline', should undergo technical validation. This includes consideration of image acquisition quality with attention to reproducibility, error and artefact which may impact on the pipeline's performance.

Clinical validation should be established prior to embedded clinical use. A useful structure for clinical validation is to perform a proof of concept *'credibility'* study followed by a clinical impact *'accuracy'* study. These concepts are further explored in section 2.1.7.

2.1.5 Step 5. Workflow integration

Integration with the established clinical workflow is essential and will increase the uptake and acceptance of the reporting tool by radiologists and referring clinicians. Basic requirements include compatibility with the data format and transfer, i.e. the Digital Imaging and Communications in Medicine (DICOM) standard, as well as integration into the PACS. Ideally the automated output of the analysis pipeline, the quantitative report, should be viewed within the same workstation environment, appearing as an additional DICOM series alongside source images. This would allow for the information contained in the report to

be efficiently integrated with the radiologist's visual assessment of the image series.

Analysis software should be developed under the appropriate quality management framework for medical devices. United Kingdom (UK) regulations are in the process of changing as a consequence of the UK leaving the European Union. Attention to patient data protection and relevant institutional information governance should be of high importance.

The National Institutes of Standards and Technology (NIST, https://www.nist.gov/), in their discussion on consideration of the role of QIBs in clinical practice, identify three central sources of uncertainty that may arise from their use. Uncertainty may arise from natural biological variability between subjects, from inconsistency of interpretation of the findings by clinical staff, and also from physical measurement variability that is associated with data collection and analysis across imaging platforms (Clarke, Sriram, and Schilling 2008). Several common challenges relating to physical measurement uncertainty are particularly important to consider for clinical QIB application, which are briefly discussed below.

2.1.5.1 Acquisition protocols

Routine clinical MRI protocols may be less sophisticated than those specified for research studies. For example, in many centres clinical T1-weighted scans may be performed with two-dimensional acquisitions. Isotropic 3D data, which is more suitable for quantitative analysis, may not be available in routine clinical practice. Inconsistencies in scanning parameters can also cause significant variation in tissue contrast. This can make, for instance, automated delineation for a subregion of interest challenging.

2.1.5.2 MRI scanner variability

Image geometric accuracy varies between scanners and vendors, resulting in varying spatial distortions which, if uncorrected, may impact upon regional tissue-volume estimates. Quantification accuracy is dependent on high reproducibility between MRI scanners. However, this is not generally a primary design concern in clinical systems, since scanner variability has a much

smaller impact on routine clinical practice based on radiologists' qualitative visual evaluation.

2.1.5.3 Image artefacts

Thorough screening of incoming data is necessary in order to detect artefacts, such as those arising from patient motion and other errors, as these artefacts may impede the automated algorithm in performing accurate quantification or produce spurious results. Adaptive correction schemes prior to analysis, such as bias field or motion artefact correction, may minimise the number of data sets failing to yield reliable volume estimates for a given measurement strategy. Many software packages are fully integrated and automated, meaning they will produce a numerical result whatever the input data and often do not allow intermediate steps to be scrutinized by the end-user.

2.1.5.4 Automation process

In order to remove reliance on time-consuming manual or semi-automated techniques requiring frequent intervention and monitoring, often by highly expert practitioners, quantification methods for clinical application should be fully automated. This also protects the process from inter-operator variability. Such automated techniques must be generalisable across the range of MRI services in the health system, including both scanner type and acquisition protocol variations.

2.1.6 Step 6. In-use evaluation

Once the automated analysis pipeline has been embedded into the clinical workflow, in-use evaluation should be undertaken with respect to the key areas of patient management and socio-economic impact. These are the key measures by which an automated quantification technique should ultimately be validated for their clinical impact and utility. Validation of automated techniques has tended to occur in the research setting largely on their technical performance alone, by comparison to results of other available methods using well curated datasets. Studies exploring the validation of these techniques once they have been embedded into clinical practice are sparse, use disparate methods and are difficult to conduct. In-use clinical validation should involve

not only testing the technical performance of the pipeline with clinical quality data, which is likely to be drawn from several different scanners depending on the particular imaging department, but just as importantly should capture the clinical interpretation and experience of the radiologist end-users. It is necessary to demonstrate a measurable benefit in terms of at least one of efficiency; interpretation accuracy; improved inter- and intra-reader reproducibility and impression confidence. Integration of quantitative reporting tools into the multidisciplinary team (MDT) setting, alongside review of a patient's clinical history and other investigations, would provide a robust setting for the assessment of whether the quantitative information was providing any added value in reaching patient management decisions. The MDT setting would allow for collection of the views of radiologists and other clinical team members on the usability of the quantitative information in clinical practice, and could help to identify any potential practical barriers that may hamper the adoption of the quantitative pipeline.

Ultimately, more widespread deployment across centres would require even further and more complex in-use evaluation strategies that include establishing standardised imaging protocols or adaptation of the analysis methods to account for site variability. The eventual aim should be for a system-wide assessment of the impact of the integrated methods on imaging services, radiologist end-users, referring clinicians and patients, as well as an economic cost-benefit assessment to the implementing hospital. This should be the long-term goal of a healthcare facility planning to implement a quantitative assessment pipeline into their imaging workflow, with a view to a long-term ongoing clinical and healthcare economic validation over the course of several years.

2.1.7 Clinical validation pathway (steps 4 and 6 of the QNI framework)

Pre-deployment clinical validation and conscientious in-use evaluation form a complete clinical validation pathway that should be undertaken when applying quantitative image analysis tools to a healthcare setting. This pathway can be summarised as relating to *credibility*, *accuracy*, *patient management*, and *socio-economic impact*.

2.1.7.1 Credibility

Validation of the proof-of-concept quantitative analysis pipeline and biological validation with real-world data should be conducted. A credibility study should take the form of a pilot study in which the chosen quantitative image analysis tool is applied to clinical MRI scans of patients with established diagnoses of the disease of interest and the results of the QIB analysis quality checked. This should include technical validation, checks on image acquisition, post-processing, analysis, and report generation.

Following these technical checks, a limited clinical validation should be performed by experienced blinded expert radiologists who have not seen the cases, and who should rate them first according to their routine practice, blinded to the quantitative report and again taking the report into account. Classical evaluation should then be compared to their impression when using the report and their consistency evaluated. Any technical or report presentation problems that are exposed at this stage should be remedied and rechecked for clinical credibility.

2.1.7.2 Accuracy

Once the credibility of the quantitative technique has been established, its impact on the clinical reporting process should be examined. This evaluation should reproduce the radiologist's normal reporting environment as closely as possible, with the quantitative report displayed alongside all relevant imaging series.

Accuracy study assessment goals should include measurement of radiologists' accuracy and may also include measurement of their subjective confidence and/or their reporting efficiency. These outcomes should be assessed both with and without the quantitative report being present.

Images should be presented in a random unpredictable order and should include a spectrum of pre-selected clinical cases, which include a mixture of disease severity from clearly pathological to more subtle changes, as well as normal-appearing control scans where available. The pathology of each case should be established to the best available gold standard, to enable

assessment accuracy to be determined. This will be specific to the condition in question, for example, cerebrospinal fluid (CSF) analysis and neuropsychiatric profile in the case of Alzheimer's Disease (AD).

The case mix should aim to reflect the spectrum and frequency of pathology encountered in a normal radiology workflow. Including a range of severity in the case mix is valuable in discerning whether the quantitative report provides added value by improving assessment accuracy where the MRI pathology is subtle. It may also be useful to include image readers with a clear range of expertise, representing the wide spectrum of training and experience levels of staff working in a radiology department. It would then be possible to establish whether the quantitative report was associated with increased inter-rater agreement between these groups, in addition to increased accuracy when assessing agreement with the gold standard. It may be of further interest to include a cohort of non-clinical image analysts in the assessment exercise, as this may uncover additional potential uses for the quantitative analysis tool, for example for consideration of its use in education and training.

2.1.7.3 Patient management

Based on the outcomes of the accuracy study, and whether this has demonstrated a measurable benefit of the quantitative report to the radiologist, the analysis pipeline should be integrated within the hospital's radiology department and the tool rolled out to specific reporting radiologists. Whether this full integration can be achieved or not will depend on technical compatibility of the tool design as well as achieving compliance with the relevant medical device regulations. This is an important stage for assessing how easily the tool can be integrated into the reporting workflow, and how the tool is performing with new clinical data, and regular structured feedback from the users should be documented.

After a period of evaluation that is restricted to named users, the tool may be more widely disseminated through the radiology department. This dissemination will require adequate staff education and training which includes both technical and clinical staff so that the department is globally capable of integrating and utilising the new technology.

To determine its impact on patient management, a prospective assessment of patients' clinical pathways, which would include assessment of the speed of diagnosis, the requirement for repeat investigations, and the basis for therapeutic decisions if available, could be compared with cases for which quantitative analysis was not available.

2.1.7.4 Socio-economic impact

Definitive socio-economic validation would require larger scale, multi-centre studies which investigate resource utilisation, productivity, clinician and patient perception, and long-term economic impact. This stage of clinical validation is particularly challenging due to the *a priori* requirement for a hospital or healthcare system to invest in such tools and their supporting infrastructure often long before an economic impact can be reliably demonstrated. Ultimately, this is the type of robust business intelligence that is required to convince purchasers within a healthcare system to bear the costs of the investment in additional software tools.

2.2 QNI framework applied to clinical neuroradiology for dementia

I will now use dementia imaging as an exemplar application to discuss the development of QIBs for clinical neuroradiology, referring directly to each step of the QNI framework that was detailed in section 2.1.

2.2.1 Step 1: Establishing a clinical need

2.2.1.1 Dementia

Dementia is a term that encompasses a class of conditions which cause irreversible progressive decline in cognitive function. It affects an increasing number of people worldwide and presents an urgent challenge for health and social care. It is estimated that more than 115 million people globally will be affected by dementia by 2050 (Winblad et al. 2016).

Underlying pathological causes of dementia are varied, and structural brain imaging can help to identify the differences between the commonest. Alzheimer's Disease (AD) accounts for 50-75% of cases, vascular dementia for 20%, and frontotemporal dementia (FTD) for 5% (Cunningham et al. 2015).

AD is characterised histopathologically by detection of neurofibrillary tangles and amyloid plaques within the brain. These cause synaptic and axonal loss and subsequent parenchymal atrophy in a progressive regional pattern (Braak and Braak 1995). This atrophy can be detected on structural imaging as affecting particular brain structures or regions in several recognised patterns or AD subtypes (Harper et al. 2014; Risacher and Saykin 2013).

Classical AD is typified by early medial temporal lobe atrophy (MTA), followed by lateral temporal, medial and lateral parietal, and frontal involvement, with relative sparing of the occipital lobe and sensory-motor cortex. This pattern on MRI is discriminating, as it is not commonly seen in normal ageing, and ante-mortem MRI findings in AD patients correlate with pathological severity post-mortem (Harper et al. 2016).

Mild cognitive impairment (MCI) is identified as a prodromal stage of AD, with a 10-15% annual conversion rate to AD, particularly within the subset with an amnestic clinical presentation (Risacher et al. 2009). Longitudinal MRI has

demonstrated that MCI subjects initially have focal, limited areas of cerebral atrophy, mainly in the medial temporal lobes, and this increases in increments to the established AD pattern (Whitwell et al. 2008). Quantification of atrophy rates, brain volume and morphometry have been useful in prediction of which MCI subjects will progress to AD, with differences from stable MCI subjects detectable well before clinical AD diagnosis (DeToledo-Morrell et al. 2004; Devanand et al. 2008; Risacher et al. 2009).

Sporadic young-onset AD differs in that it is more likely to be non-amnestic, and structural imaging most commonly demonstrates biparietal or bi-parieto-occipital atrophy, termed posterior cortical atrophy (PCA) (Rossor et al. 2010). This pattern is not pathognomonic and can also be seen in dementia with Lewy Bodies (DLB), corticobasal degeneration (CBD) and prion disease (Crutch et al. 2012). A recent classification system aims to define the PCA syndrome based on clinical and radiological features, highlighting the importance of structural imaging in this setting (Crutch et al. 2017).

2.2.1.2 Structural MRI for dementia

Structural MRI is the mainstay of conventional neuroradiology in current dementia practice (Wattjes 2011). In the diagnostic setting, it is important for the exclusion of alternative pathologies, and for establishing the regional pattern of atrophy to differentiate between the dementia pathologies themselves (Staffaroni et al. 2017). A subjective estimation of disease severity or stage can be made by the radiologist based on the degree of atrophy present in the context of what would be normal for the patient's age.

It is not only challenging to establish an early clinical diagnosis of dementia, but moreover, the first signs and symptoms can emerge years if not decades after the underlying pathophysiological process have been initiated (Counts et al. 2016). Recently therefore, dementia research has prioritised the identification of potential clinical and imaging biomarkers early in the disease course (Risacher and Saykin 2013).

Validated and clinically adopted QIBs could facilitate diagnosis in the early or even prodromal disease phases; provide objective measures of difference from the normal aging spectrum; exclude differential diagnoses; and support

powerful preclinical drug trials (McEvoy and Brewer 2010a; Salvatore, Cerasa, and Castiglioni 2018). There is also a potential role for QIBs as prognostic measures, since the relationship between progressive biomarker change and clinical disease trajectory can be modelled (Caroli and Frisoni 2009).

With imaging identified as a potentially powerful diagnostic tool, the introduction of objective methods for quantifying dementia-related changes on MRI, especially MTA assessment in AD, has become a priority for clinicians and researchers. Indeed, the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association, NINCDS-ADRDA, issued updated guidelines which incorporated structural MRI of the medial temporal lobe in the assessment of AD; however, no specific method for objective measurement was defined (Cedazo-Minguez et al. 2016).

2.2.1.3 Current practice: visual rating scores

Visual rating scores, with ‘cut-offs’ defining degrees of abnormality, are a semi-quantitative means aimed to facilitate communication among practitioners. MTA grading was established by Scheltens et al. using a discrete 5-point scale for the size of the hippocampal formation as well as the prominence of adjacent cerebrospinal fluid (CSF) spaces (Scheltens et al. 1992). Whilst this rating system has proven useful for rapid screening in clinical practice, and has good inter-observer agreement (Boutet et al. 2012), the ratings by definition remain subjective and reliant on the radiologist’s experience. The cut-off for abnormality varies by age (Pereira et al. 2014) and is somewhat arbitrary.

Other visual rating scores were also developed to characterise additional or atypical features that can be seen in AD. The Koedam scale focuses on features of PCA, showing good inter-observer agreement and an 87% specificity for AD when used in combination with MTA scoring (Koedam et al. 2011).

Visual rating scores are useful, rapid and accessible tools in clinical practice. Their limitations include insensitivity to subtle or early changes, ambiguity in distinguishing pathological change from normal ageing, ceiling and/or flooring effects, and being only coarsely discriminating due to their discrete

categorisations. They are used to varying degrees by radiologists across Europe, depending on levels of training and with unknown reproducibility both within and across imaging departments (Vernooij et al. 2019). Fully quantitative imaging biomarkers may largely address these issues providing practitioner-independent objectivity, at least across a single radiology service.

2.2.2 Step 2: Developing a method for automated quantitative analysis

Many biomarkers show promise in early research phases but face bottlenecks when it comes to validation and clinical implementation. Frisoni et al. proposed a 5-phase framework for development of specific QIBs for prodromal AD, adapted from a framework implemented in oncology screening (Frisoni et al. 2017). Despite being one of the most established imaging biomarkers in AD, MTA is still in the early stages of this five-step pathway (ten Kate et al. 2017). Standardised validation procedures for automated segmentation algorithms based on a harmonised manual segmentation protocol is identified as an imminent priority, which will allow for meaningful assessment of algorithm reproducibility. Only then can their clinical validity and utility be evaluated in the memory clinic. The Frisoni framework integrates well with the QNI framework in that it focuses on QIB development and clinical validation, which are key parts of the broader, end-to-end translational implementation pathway detailed by the QNI.

In the research context, numerous studies have compared cerebral atrophy between AD, MCI and healthy control groups using MRI-based regional volume measurement (“volumetric”) approaches rather than visual rating scores. Measurement methods have included manual delineation of anatomical regions of interest, as well as automated or semi-automated volumetry, although with varying protocols for anatomical delineation for segmentation (Caroli and Frisoni 2009). More recent methods allow for the parcellation of brain components into grey matter (GM) white matter (WM) and CSF and for automated voxel level quantification (Despotović, Goossens, and Philips 2015; Matsuda 2016). Methods for detecting and quantifying white matter vascular disease burden and mitigating the effects of severe WM damage on the success of the volume quantification algorithm, will be

particularly important in the dementia and ageing populations (Chard et al. 2010).

2.2.3 Step 3: Communicating quantitative analysis results

Presenting the outputs of quantitative analysis tools in a clinically meaningful way is key to their translational success. Several commercial software solutions for clinically applied global and regional brain analysis are already available and in use, the mainstay of which are for dementia imaging but also for use in other neurological conditions such as stroke and multiple sclerosis. There is growing clinical interest in these tools, which have undergone various degrees of technical and clinical validation. As previously mentioned, NeuroQuant is a software package based on the Freesurfer algorithm with a large normative reference database offered by the company cortechs.ai (www.cortechslabs.com, cortechs.ai 2021). Its report includes coloured overlays of the segmentation output, tabulated regional brain volume results, and normative reference graphs where the subject's results are plotted on a graph displaying percentiles of normative reference volumes. It has been used in clinical validation studies of AD (Min et al. 2017; Persson et al. 2018), where its results were compared to visual assessment alone, as well as in other disease areas such as traumatic brain injury (Ross et al. 2015) and temporal lobe epilepsy (Azab et al. 2015; Louis et al. 2020).

It is however important to recognise that although commercially available tools have the appropriate regulatory marking (e.g. CE in the European Economic Area, or United States Food and Drug Administration, FDA,) this does not necessarily mean that a solution has been fully validated in terms of its clinical impact and validity. Regulatory priorities are to demonstrate that the solution reliably produces reproducible results and is therefore no direct risk of harm to patients, so the process largely prioritises documentation and framework implementation for device deployment. Clinical import or efficacy, which are key parts of the QNI framework, are not the focus of regulatory approval and therefore in isolation CE marking may provide false reassurance regarding the appropriateness of introducing a product into clinical use (Feldman et al. 2008; Mishra 2017).

Appropriate reference data for contextualisation of individual subject quantitative results is key to QIB clinical interpretability. Reference data from healthy control subjects with comparable acquisition parameters that cover the appropriate clinical age range should be processed using the same analysis pipeline. Standardisation of acquisition protocols for dementia is being tackled in part by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jack et al. 2008), who provide a rich source of multi-centre data. Their suggested protocols are increasingly being adopted by clinical institutions, which will enable reference data to become more widely applicable between centres.

2.2.4 Step 4: Pipeline validation

The validation of brain segmentation methods is at best challenging, due to the general absence in general of ground truth data: most validation studies have been based on cross-validation with the performance of alternative automated techniques, using data from the same source and with the same restrictions (e.g. single-site and vendor, precise scanning parameters). For repurposing research-developed QIBs into robust neuroradiological tools, it is important to note that most QIB research has focussed on group-level discrimination, which does not necessarily mean that the technique under consideration will offer sufficient sensitivity to support single-subject classification inference (Brewer 2009; Chard et al. 2010). Another challenge is that much of the available data used for the development and validation of QIBs in controlled research studies may poorly represent relevant clinical populations by neglecting comorbidities, socioeconomic status and education (Arbabshirani et al. 2017), i.e. the real-world scenario of patients with neurological comorbidities and other structural brain abnormalities.

2.2.5 Step 5: Workflow integration

A software platform including image data identification and routing functionality is required to support integration of the QIB analysis and report generation into the hospital electronic information systems, including the PACS. A means of identifying examination images series appropriate for analysis is required, in our case implanted using customised DICOM series labels. Careful consideration should be given to the problem of clinical case stratification, so

that reports are generated and interpreted in the appropriate context of conventional imaging and clinical history, and unwarranted quantitative reports are not generated. Introduction of QIBs into clinical dementia reporting may require modifications to the referral process. DICOM image tags could be queried by the quantification module to ensure that the report is generated for the correct compliant series.

2.2.6 Step 6: In-use evaluation

This should initially occur at a departmental level, ideally with engagement overseen by a key senior neuroradiologist. Initial training and technical systematic evaluation should be followed by a small pilot evaluation involving experienced radiologists who are familiar with dementia imaging reporting. They should report as per their normal routine, and by comparison corroborate the consistency and reliability of the quantitative report. Feedback should be gathered on any discrepancy or technical difficulty encountered. Following this pilot, all reporting radiologists in the department would be expected to adopt the report as part of their dementia imaging assessment. This prerequisites training and engagement not only of the radiologist team but also of the referring clinical teams. Audit of reporting efficiency and patient management pathway timelines would provide measures of service impact in comparison to previous practice. Ultimately, higher-level in-use evaluation of the outcome benefits of quantitative reporting will require larger-scale, multi-centre studies once adoption by single centres has become well established.

2.3 Conclusions

The QNI framework seeks to address the many and varied challenges that exist to enable the translation of QIB reporting to the clinical setting. There remain some fundamental obstacles to translating these promising imaging biomarkers into clinical practice, and the QNI framework provides a structured technical and clinical validation process to address these. Lack of large-scale and rigorous technical and clinical validation, and over-reliance on CE marking for quick commercial deployment, are major potential pitfalls in the field. The greatest challenge may be the circular problem of establishing clear evidence of clinical and socio-economic benefit, without prior wide-scale adoption. This is especially challenging for quantification of conditions that do not currently have disease-modifying treatments available, as ground truth is not known, and other indicators must be relied on when assessing the added value of a diagnostic aid.

Using dementia as an exemplar for QIB translation, it is clear that imaging and clinical biomarkers are adapting our perception of dementia from a largely clinical and post-mortem diagnosis of exclusion to a more systematic biological paradigm. Incorporating QIBs for dementia into clinical reporting workflows via frameworks such as QNI may support earlier and more certain diagnosis. It is important to balance the potential that these imaging biomarkers hold with the realisation that, within the dementia field, they are still in the early stages of validation. Expediting translational pathways, for both dementia and other important neuroradiological indications, will require consensus on prioritisation of key issues including protocol harmonisation, data sharing, and algorithm development.

In the following chapters, I will apply aspects of the QNI framework to translational QIB development in four disease areas, with a focus on technical and clinical validation: hippocampal sclerosis, dementia, multiple sclerosis, and glioma.

3. Quantitative reporting for hippocampal sclerosis

3.1 Introduction

3.1.1 Epilepsy and hippocampal sclerosis

Epilepsy is a neurological disease affecting approximately 0.5-1% of the global population (Choi et al. 2008) characterised by the abnormal discharge of neurological activity in the brain. When this abnormal electrical activity occurs, an individual will experience irregular neurological functioning, which may be manifested in a wide variety of forms, from typical seizures to abnormal sensations, to periods of absence or loss of awareness. Epilepsy can be broadly categorised as generalised or focal; generalised epilepsy involves abnormal electrical activity across the whole brain and is associated with impaired consciousness, whereas focal epilepsy arises from a specific location in the brain and may or may not affect consciousness. Most of the one third of people who suffer from pharmacoresistant disease have a focal epilepsy (Sidhu, Duncan, and Sander 2018).

Focal epilepsies often have an identifiable underlying structural cause that can be identified on MRI. Causes include focal cortical dysplasia (FCD), vascular malformations and hippocampal sclerosis (HS). HS is the most common underlying cause of mesial temporal lobe epilepsy, which itself is the commonest form of epilepsy in humans (Engel J. 2001). Although the pathogenesis of HS is still unclear, there is a strong association with early childhood precursor injuries, which are most commonly childhood febrile seizures but also include birth trauma and head injuries (Mathern, Pretorius, and Babb 1995). Seizures arising from HS pathology are more often resistant to pharmacological therapies than other epilepsy types, and surgical excision of the epileptogenic focus can be curative (Engel et al. 2003).

3.1.2 Histological and neuroimaging biomarkers of HS

HS is characterised histopathologically by neuronal loss and tissue gliosis (Thom et al. 2009) affecting the pyramidal cell layer of the cornu ammonis (CA) (Coras and Blümcke 2015). These pathological hallmarks are identifiable by their representative features on structural MRI scans, which appear as hippocampal volume loss (atrophy), disruption of normal hippocampal

architecture, and increased T2-weighted signal intensity (Briellmann et al. 2002; Van Paesschen 2004). Hippocampal atrophy secondary to the damage and loss of neuronal cells can be identified radiologically by loss of height in the hippocampal formation itself in combination with widening of the cerebrospinal fluid (CSF) spaces that surround it, these being the choroid fissure and temporal horn of the lateral ventricle. Histopathological sclerosis of the hippocampus is characterised by replacement of neuronal cells with astrocytes and microglia, a pathological process known as gliosis. These changes of cell type alter the water content and therefore the T2 values of a voxel, and can be detected as an increase in T2 relaxation time (Peixoto-Santos et al. 2017).

3.1.3 Identifying HS imaging biomarkers on MRI

A commonly accepted standard protocol for structural MRI in patients with epilepsy aims to achieve diagnostic sensitivity to possible aetiologies while maintaining clinical practicality (Wellmer et al. 2013):

a) Three-dimensional isotropic volumetric T1-weighted series for detection of cortical malformations or hippocampal volume loss;
b) Axial and coronal T2-weighted series to assess hippocampal architecture;
c) Fluid attenuated inversion recovery (FLAIR) series, useful for detection of HS and FCD; and
d) Axial T2* gradient echo (GE) or susceptibility weighted images (SWI), used to detect vascular or calcified epileptogenic lesions.

Identification of features consistent with HS on MRI is used in conjunction with other clinical investigations to make a diagnosis of HS as the cause of a patient's epilepsy. Moreover, accurate localisation of these HS imaging features will enable surgical planning to guide the excision of the epileptogenic focus, the aim being to achieve complete resolution of seizures and spare as much healthy tissue as possible to limit post-surgical morbidity (Duncan and Sagar 1987; Lencz et al. 1992).

While accurate interpretation of these imaging features is straightforward when the hippocampus is significantly atrophied and gliosed, these signs are often

more subtle and equivocal earlier in the disease, and accurate detection becomes dependent on the experience and subjective opinion of the reporting radiologist. There is a particular challenge in accurately detecting bilateral cases of HS, as their symmetrical appearance can obscure the presence of pathology. Asymmetrical bilateral HS, where one hippocampus is more sclerosed or atrophied than the other, could potentially lead to an erroneous diagnosis of unilateral HS. Detection of bilateral HS is vital to avoid performing a non-curative surgical resection. Other challenges that impair detection of HS include where the patient's head is positioned asymmetrically so coronal comparison between sides is difficult without reformatting, and importantly if there is concomitant parenchymal volume loss consistent with ageing hippocampal atrophy that is due to HS pathology will be less prominent.

A visual rating study assessing inter-rater agreement on MRI detection of hippocampal volume loss in unilateral HS indicated a threshold effect where the abnormality was only detected at an asymmetry ratio of 0.7 or lower compared to the unaffected side (Reutens et al. 1996). This suggests that more subtle volume loss or unrelated atrophy affecting the contralateral hippocampus would cause the true pathology to be missed. In a similar vein, raised T2-weighted signal can also be very difficult to detect, particularly in bilateral HS, not least because the hippocampus and its neighbouring structures in the limbic lobe possess innately higher T2-weighted signal than the rest of the brain (Asao et al. 2008; Hirai et al. 2000). HS can be characterised on MRI by T2-weighted abnormality alone, with no significant volume loss (Meiners et al. 1994). Therefore, for cases of subtle and/or bilateral HS, accurate diagnosis based on subjective assessment of structural MRI is extremely challenging.

3.1.4 Quantification of HS neuroimaging biomarkers

Quantitative analysis of hippocampal volume and T2-weighted signal properties has been shown to reduce subjectivity and assist the radiological assessment of suspected HS (Bernasconi et al. 2000; Van Paesschen 2004). Volumetric analysis performed by segmenting the hippocampus (which was

initially done manually and has more recently been automated) has been shown to increase diagnostic sensitivity (Coan et al. 2014; Martins et al. 2016).

T2 relaxometry, also referred to as quantitative T2 (qT2), describes the quantification of a tissue's T2 relaxation properties. T2 relaxometry has been shown to increase detection of HS (Jackson et al. 1993), especially given that T2 signal change may be the only sign of HS with no significant volume loss (Bernasconi et al. 2000). Quantification of T2 hyperintensity allows for objective assessment of one of the key radiological features of HS, and has been shown to be more sensitive than visual inspection (Namer et al. 1998; Van Paesschen et al. 1995). The technique has also been automated and been shown to have higher reproducibility than manual processing (Winston et al. 2017). Quantification of FLAIR signal is also possible (Huppertz et al. 2011), but found not to be as powerful as qT2 (Rodionov et al. 2015).

3.1.5 Aims and hypothesis

Based on the framework for imaging biomarker validation and clinical translation that were set out in Chapter 2, the work in this chapter is focused on applying this translational framework to technical and clinical validation of quantitative methods for analysis of hippocampal atrophy and sclerosis in individuals with HS.

With the knowledge that these imaging biomarkers have been automated and shown to improve HS detection, the aim of this work is to utilise these quantitative techniques in an accessible and clinically meaningful way, by validating a visually meaningful quantitative report for application to an individual subject that provides comparison to a normative reference population. It is hypothesised that the use of such a report will increase detection of visually equivocal cases of HS and reduce disagreement between image interpreters, both of which will enhance the clinical decision-making process in the management of patients with HS.

3.2 Methods

3.2.1 Hippocampal segmentation

Segmentation of the hippocampi was automatically performed using a previously described publicly available (https://hipposeg.cs.ucl.ac.uk/) method (Winston et al. 2013). The T1-weighted image is segmented using a multi-atlas-based segmentation algorithm (Similarity and Truth Estimation for Propagated Segmentations, 'STEPS', (Cardoso et al. 2013)) based on a template database of 400 manual segmentations complied from clinical epilepsy scans, which includes a wide spectrum of HS severity as well as other epileptogenic pathologies. STEPS builds on a propagation method (simultaneous truth and performance level estimation, 'STAPLE', (Warfield, Zou, and Wells 2004)) and adapts it to perform label fusion based on local similarities for template selection.

Group-wise templates of the hippocampal segmentations are created by first registering each image to an arbitrarily selected reference image using affine transformation to compute an average image, followed by repeated iterations of registering all images to the average image using nonrigid transformations, which is repeated until convergence is reached.

The individual subject scan to be segmented is nonlinearly registered to the group template and a region of interest (ROI) extracted from the subject scan based on the group hippocampal segmentation templates. Following this each template scan is coarsely registered to the subject's ROI and the 75 most closely correlated subjects are selected for a second more accurate registration. Finally, the 15 most similar cases at each voxel are selected and fused using a probabilistic framework to iteratively estimate the genuine segmentation. Hippocampal volumes are corrected for intracranial volume by applying the same algorithm for whole brain segmentations.

There are several strengths to this multi-atlas based approach. Use of a method which employs an atlas of multiple template images is superior to using a single template (Heckemann et al. 2006). Most other atlas-based segmentation methods have used small template datasets comprising healthy subjects. Comparison of segmentation performance has shown that results for

TLE are much worse than for healthy controls or subjects with Alzheimer's Disease (AD) (Kim et al. 2012). The presence of temporal lobe epilepsy is associated with a high level of atypical hippocampal configuration or location (Bernasconi et al. 2005). Use of this large multi-template atlas with a variety of hippocampal pathologies performs better for segmentation of HS scans, which have different morphology to hippocampal pathology due to AD (Duan et al. 2020).

3.2.2 T2 relaxometry

T2 relaxometry, also referred to as quantitative T2 (qT2), measures the intrinsic T2 relaxation time of a specific tissue. Automated T2 relaxometry maps were calculated for each voxel from the T2 signal at two different echo times using a monoexponential fit as described and technically validated by Winston and colleagues (Winston et al. 2017).

Once the hippocampi have been segmented using the STEPS algorithm described in section 3.2.1, voxelwise T2 maps are calculated from the signal S_1, S_2 at two echo times TE_1 and TE_2 using a monoexponential fit:

$$T2 = \frac{TE_2 - TE_1}{\ln\left(\frac{S_1}{S_2}\right)}$$

The T1 hippocampal segmentations are registered to the dual-echo PD/T2 image and then eroded. Cerebrospinal fluid (CSF) signal contamination is minimised by eliminating voxels with T2 values higher than 170 ms. The mean T2 value can then be calculated in the defined ROI. The CSF cut-off value of 170 ms was determined by measuring T2 values in a healthy control population within each brain compartment (grey matter, white matter, and CSF) and finding that 170 ms was the optimal threshold between higher grey matter T2 values and lower CSF values with the minimum overlap.

This method has been shown to reliably separate healthy controls from subjects with HS, have good agreement with manual methods and in fact have superior reproducibility on repeated measurements (Winston et al. 2017). Automation replaces the onerous manual approach of delineating ROIs which are by necessity limited to a fraction of the real hippocampal volume in order

to avoid CSF contamination. By using automated segmentation as the template, mean T2 values can be measured across the whole hippocampus. Using an erosion at the boundary and thresholding of higher T2 values, the issue of CSF contamination is minimised.

Combined assessment of hippocampal volume and qT2 is both sensitive and specific for HS (Coan et al. 2014).

3.2.3 Cross-sectional volume and qT2 measurements

The methods described above generate single global values for automated volumetry and mean qT2 measurement across the hippocampus. HS may be a diffuse or focal process, meaning that these techniques may lack sensitivity to subtle focal pathology (Woermann et al. 1998). The ability to quantify these biomarkers in a single subject at a more localised level and build a profile along the length of the hippocampus would address this limitation, and a method to achieve this has recently been proposed (Vos et al. 2019). It uses the global processing methods that have already been described (Winston et al. 2013, 2017) and extends them to obtain profiles along the long axis (anterior-posterior, A-P) of the hippocampus.

This methodological extension involves the creation of a normative database to establish group average values, against which the single subject's scan can be compared. Briefly, this process involves the generation of a group template where a reference dataset of healthy control subjects' scans is registered to an atlas of the Montreal Neurological Institute (MNI) template that has been reorientated to place the hippocampal ROIs along the A-P axis. This was done with 111 healthy control subjects to generate a dataset-specific group average template. The template can then be assigned a distance map along the A-P axis from the most posterior (1mm) to most anterior (218mm) slice. Hippocampal segmentations generated for the reference dataset are transformed to this group template.

An individual subject's T1 scan is registered to the group template and for each coronal slice the distance map that was created for the group is transformed to the subject's scan, to map the A-P locations for the individual subject in relation to the template. Hippocampal cross-sectional area (CSA) is then

calculated for the subject's segmentation, corrected for total intracranial volume (TIV).

The transformation from the template to a subject's T1 scan is concatenated with a transformation from the T1 to PD/T2 scan for the T2 relaxometry processing. Template A-P location mapping is similarly transformed to the hippocampal segmentation in PD/T2 space to obtain location information for the individual subject.

In order to contextualise the cross-sectional quantification of hippocampal volume and qT2 for an individual subject, the reference population's corresponding values have been used to generate normative reference ranges. At every 1mm along the A-P hippocampal axis, the mean and standard deviation of CSA and qT2 was calculated from the dataset of 111 healthy controls. Normative reference ranges (mean±1.96 x standard deviation) were then calculated for CSA and qT2 for each location along the long axis of the hippocampus.

To visualise this information, plots along the A-P axis of the hippocampus displaying reference ranges for both CSA and qT2 were generated. As CSA is derived from the isotropic T1 scan, the individual subject's profile is shown as a continuous line overlayed on the reference plot. As qT2 values are derived from a 2D PD/T2 acquisition, with 4mm slice thickness, individual subject data points are represented discretely for each slice.

3.2.4 Quantitative report design

These methods allow for the construction of a quantitative report of hippocampal volume and qT2 values, both as global values and as profile values along the long axis of the hippocampus, to enhance the detection of focal pathology. The report that has been generated includes global CSA and qT2 results in a main table, presented with global refence ranges, as well as left to right CSA and qT2 ratios. Additionally, the report continues with a series of profile plots depicting left and right hippocampal CSA and qT2 along the P-A axis, with the reference data range shown in purple, and the individual subject's values overlayed. Left and right A-P plots are also presented overlayed on one plot to provide a visual representation of any possible

asymmetry between the two sides. The report also includes snapshots of the hippocampal segmentations, so that these can be rapidly checked for any gross errors. Subject's demographic information including age, gender, scanner, and scan date are also populated to complete the report (Figure 3-1).

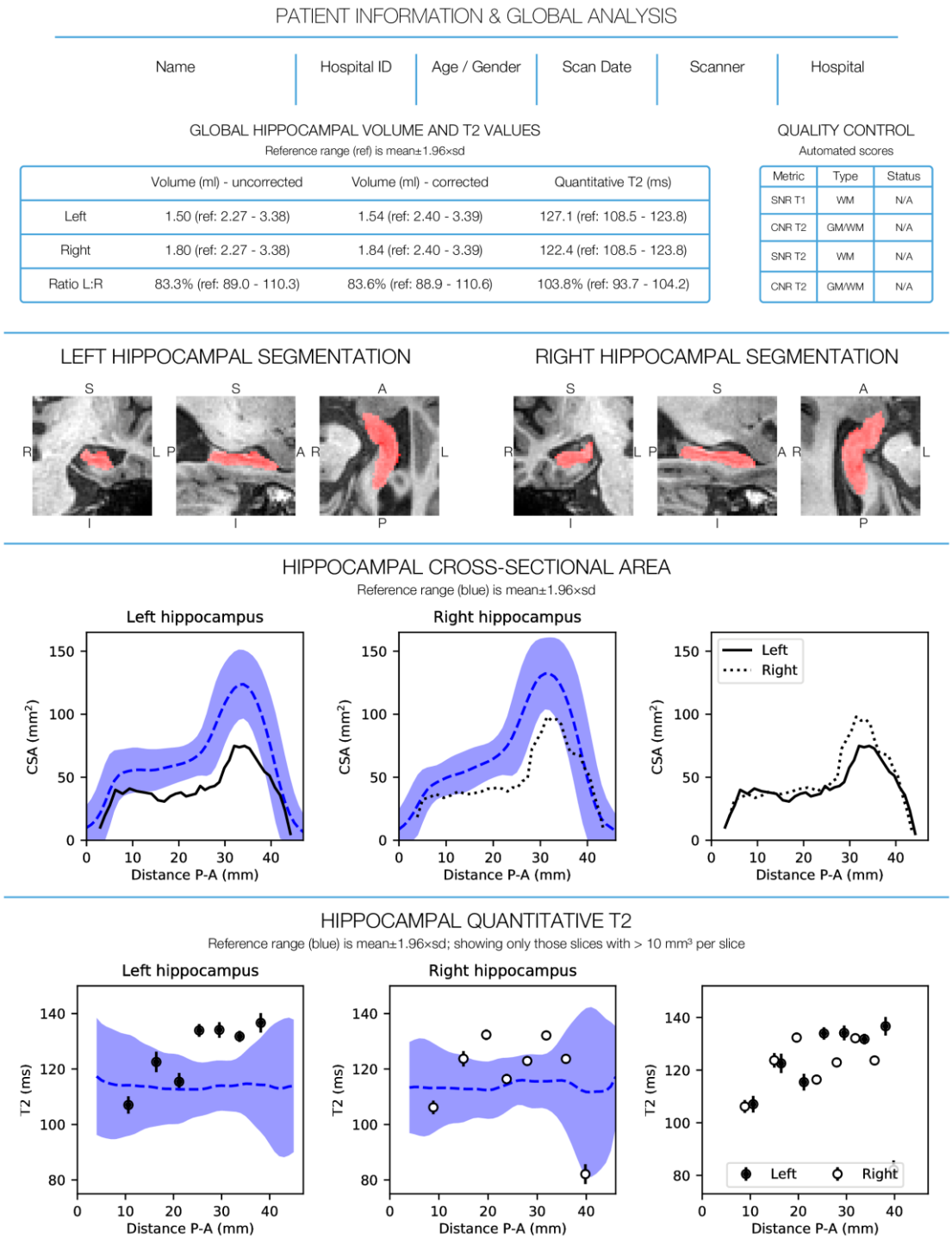


Figure 3-1. A quantitative report for a subject with bilateral HS.

3.3 Credibility Study

3.3.1 Aims of the credibility study

The aims of performing a credibility study were to assess the reliability of the quantification pipeline which populates the final HS report when applied to clinical MRI studies. The performance of the automated techniques that feed into the report were assessed, including a review of the accuracy of the automated hippocampal segmentations, and the credibility of the numerical values and graphical representations for hippocampal volume and T2 relaxometry. Results were reviewed to assess if they are adequately expressed in the context of normative data. Importantly, review of whether the report reflects the clinical impression made independently by expert radiologists was performed.

3.3.2 Methods

Ten subjects with a radiological diagnosis of HS and 10 healthy controls were selected. Subjects had undergone imaging on a 3T GE MR750 scanner with a 32-channel coil. Sequences included a three-dimensional (3D) T1-weighted inversion-recovery fast spoiled gradient recalled echo, a 3D T2 Fluid-Attenuation Inversion Recovery (T2-FLAIR), and coronal dual-echo fast recovery fast spin echo proton-density/T2- weighted.

Quantitative reports were generated for each of the twenty subjects using the methods described in section 3.2. Two expert neuroradiologists assessed the T1-weighted and FLAIR images for each subject and made their clinical impression (HS right, left, bilateral; or normal). They also evaluated the accuracy of the hippocampal segmentations by viewing them as overlays on the T1 images using the image viewing platform niftiMIDAS. Finally, they appraised the QNI report for whether it was reflective of their clinical impressions. The reports were also assessed for any technical or presentation issues.

3.3.3 Results

3.3.3.1 Overall report credibility

Rater 1 found that 17 QNI reports were consistent with their clinical impressions. For the three that were not consistent, the common issue affecting these cases was that the hippocampus appeared pathologically hyperintense on FLAIR, but the qT2 value appeared in normal range and was not reflective of the expert's impression of the presence of pathology.

Rater 2 found that 18 reports were consistent with their clinical impressions. For the two that were not consistent, they noted one had isolated points above normal range in the qT2 value, where the FLAIR appeared normal intensity, and the other was a case where the hippocampal volumes were reported as more asymmetrical than expected from the expert's visual impression.

3.3.3.2 Hippocampal segmentations

Rater 1 approved all forty hippocampal segmentations as credible.

Rater 2 identified 11 of the forty segmentations as having slight errors, eight of which were identified as under-segmentation of the hippocampal tail. These errors were described as minor and did not affect the expert's overall impression of the credibility of the QNI report. The manual delineations in the template database of Winston et al. 2013 were delineated using the guidelines described by Cook et al (Cook et al. 1992), which may vary from other delineations (Konrad et al. 2009).

3.3.3.3 Report layout and presentation

There were a few minor issues identified in the report presentation. In one case the CSA volume plot for a patient who had a high hippocampal volume that reached outside the normative range was cropped out of the graph. Another report demonstrated that the legend for the qT2 asymmetry graph was encroaching on the data points and appearing as part of the results. A minor observation was that the standard error bars for the qT2 values were difficult to appreciate as they were in blue and blended with the shaded blue of the normative range.

3.3.4 Outcomes

Where the qT2 values were within normal range but the FLAIR was hyperintense, it was discovered that some of the qT2 values for these very sclerotic hippocampi were so high that they were exceeding the threshold that is used to exclude CSF signal. This limitation must be considered and highlights that the report should be viewed alongside the images for standard visual assessment. A limitation of the current report is that no ‘sense check’ is given to report reader to indicate that this error has occurred, which could be addressed in a user guide for a future report version. Technical solutions to this issue could also be explored for example by more stringent erosion of the hippocampal boundaries.

Comments from rater 2 on the minor segmentation errors were mostly concerning under-segmentation of the hippocampal tail. The boundaries used were based on a template dataset of 400 segmentations, which followed a certain definition of anatomical boundaries. The exact anatomical border of the hippocampal tail cannot be identified on imaging alone and therefore a somewhat arbitrary definition must be imposed. The rater was satisfied that the segmentation technique did not affect the credibility of any of the QNI reports.

The small presentation issues were also addressed. The issue of premature cut-off was amended by making the maximum threshold of the y-axis scale depend on either the patient maximum or the normative curve maximum, whichever is higher. The qT2 asymmetry legend was amended by clearly separating it from the subject’s data points. The qT2 standard error bars were changed from blue to black.

Performing this credibility study has shown the HS QNI reports to be wholly credible in representing hippocampal volume and quantitative T2 relaxometry measurements. Issues have been addressed where they were identified, as described above.

3.4 Clinical accuracy validation study

3.4.1 Aims and hypotheses

As set out in chapters 1 and 2, a crucial step along the quantitative imaging biomarker translational pathway is robust pre-use clinical validation of the analysis tool. It was designed to assess whether the quantitative report, deemed credible by experts, impacted of the detection accuracy and/or confidence of image reporters of several different levels of prior expertise. Performing this multi-rater comparator study allowed several hypotheses to be tested: that by using this quantitative report for HS inter-rater variability will decrease; HS detection and rater confidence will increase; and that there will be an identifiable effect across the groups of raters, with most benefit seen to the least experiences image readers.

3.4.2 Methods

The clinically acquired data used was considered a service improvement by the National Hospital for Neurology and Neurosurgery and the UCL Queen Square Institute of Neurology Joint Research Ethics Committee. Informed written consent was obtained from control subjects. Study subjects have previously been utilised in Vos et al. 2019.

3.4.2.1 Test dataset

Forty-three subjects who had been scanned on the same 3T GE MR750 scanner with 32-channel coil were included, mean age±SD 40.0±14.8 years, 22 males. Twenty of these subjects were patients with HS and 23 were age-matched patients with epilepsy who had no focal abnormalities on MRI (referred to as ‘MR negative’ epilepsy). Of the 20 HS patients, 15 had a histologically-confirmed unilateral HS and 5 had bilateral HS which was determined by a consensus of MRI, neurophysiology and semiology.

3.4.2.2 Reference dataset

A reference dataset of 111 healthy control subjects underwent imaging with the same scanner and scanning protocol, age±SD 40.0±12.8 years, range 17.0-66.6 years, 52 males), as part of a previous study (Vos et al. 2019).

3.4.2.3 Imaging protocol

The imaging protocol comprised:

a) 3D T1-weighted inversion recovery fast spoiled gradient recalled echo (3D T1); field-of-view (FOV): 224×256×256 mm (antero-posterior, left-right, inferior-superior), acquisition matrix: 224×256×256, voxel size 1 mm isotropic, echo/repetition/inversion time (TE/TR/TI) = 3.1/7.4/400 ms; flip angle 11°; parallel imaging acceleration factor 2;
b) 3D T2-weighted fluid-attenuation inversion recovery (T2-FLAIR) sequence; a 3D fast spin echo (FSE) sequence with variable flip-angle readout (CUBE); FOV, matrix, and angulation identical to the 3D-T1, but with TE/TR/TI = 137/6200/1882 ms (Vos et al. 2018);
c) coronal dual-contrast fast recovery fast spin echo proton-density/T2-weighted (PD/T2) sequence for T2-quantification; FOV: 220×220, matrix: 512×512, in-plane resolution: 0.43x0.43mm, 55 slices of 4 mm thickness (TE effective 30 and 119 ms, TR 7600 ms, SENSE factor 2).

3.4.2.4 Assessment task

Nine raters were invited to participate in the assessment exercise, where they assessed the test dataset images and recorded their impressions both with and without the availability of the quantitative report. Reports were presented to the raters in a fully randomised manner. The nine raters made up three groups of differing radiological experience; three were expert raters who were established consultant neuroradiologists; three were radiologists completing their specialty training in neuroradiology; and three were image analysts, a group made up of MRI radiographers who worked in neurology centres and non-clinical research fellows working in epilepsy.

The raters were blinded to diagnosis and asked to review each of the 43 cases twice, in a randomly generated order, once with and once without the quantitative report available. Raters were asked to select whether a case was normal or abnormal, give their confidence from 1 (no confidence) to 5 (full confidence). If the rater selected that a case was abnormal, they had to make

a further selection of whether they thought the case was right, left or bilateral HS, and again give their confidence level in this decision using the same 1-5 scale. The only demographic information presented was the subject's age and gender.

3.4.2.5 Assessment platform design

The assessment exercise was hosted on a website designed for this purpose, to enable raters from multiple centres and geographical locations to participate. It also ensured that a standardised assessment platform was being used by all raters.

The website platform that raters used to perform their assessments gave an initial introduction to the project and a description of the task, as well as an explanation of the layout and contents of the quantitative report, and instructions on how to use the website. Once a rater was ready to begin the assessment, the website presented the cases to them in a predefined randomly generated manner, so that the rater would at separate and unpredicted points in the exercise encounter one case twice, once with and once without the quantitative report alongside the image series (Figure 3-2).

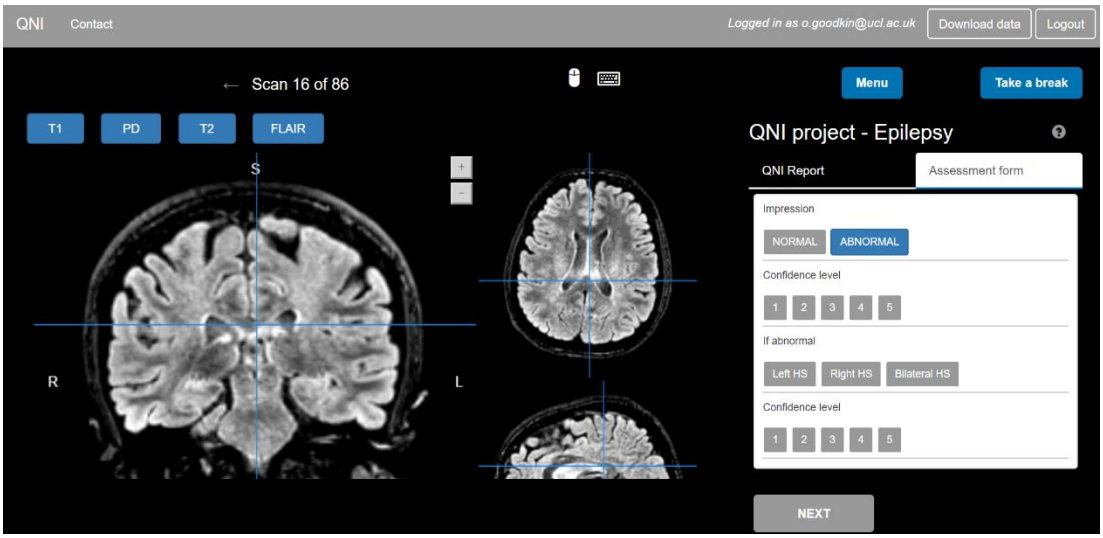


Figure 3-2. A snapshot of the website platform where raters performed the assessment task. T1, PD, T2 and FLAIR sequences were available in interchangeable panels. The assessment form was tabbed alongside the quantitative report.

The MR studies were presented in three interchangeable orthogonal planes (sagittal, coronal and axial), and raters could toggle between the four available

series (T1, proton density (PD), T2 and FLAIR) in order to imitate the radiologist's normal working environment as closely as possible.

3.4.2.6 Statistical analysis

To determine the effect of the quantitative report on diagnostic accuracy, tests from signal detection theory were used (REF). In comparison to the gold standard diagnosis, rater assessments were assigned as:

a) True positive (TP) – correctly abnormal
b) True negative (TN) – correctly normal
c) False positive (FP) – incorrectly abnormal
d) False negative (FN) – incorrectly normal.

Accuracy was therefore defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} x100$$

Data analysis was hierarchical, starting with counts of correct and incorrect assessments as normal or abnormal, against the clinicopathological gold standard. Resulting counts with and without the quantitative reports was assessed with a McNemar test. Paired t-tests were used to analyse mean accuracy and sensitivity between the report vs no report settings. Cohen's d effect size was used to assess the standardised differences in means, with a level of d>0.8 being defined as a large effect size (Cohen 2013).

Agreement of each rater with the gold standard was assessed using Cohen's Kappa, a metric which accounts for chance agreement (Cohen 1960). Agreement can be defined as moderate with a Kappa score of 0.60-0.79 and strong with a value of 0.80-0.90 (McHugh 2012). The Kappa scores were then compared using paired t-tests between the report vs no report settings. The analysis steps were then repeated for correct and incorrect rater lateralisation of right, left or bilateral HS with and without the quantitative report. Agreement between raters and reliability were assessed with Cronbach's alpha and intra-class correlation (ICC).

Mean confidence ratings with and without the quantitative report were assessed using paired t-tests. Effect size was measured with Hedges' g_z (Durlak 2009). In a further exploratory analysis, mean confidence scores per rater were split by a) whether the diagnosis made was correct or incorrect, b) whether or not the quantitative report was available and c) rater experience level. This was done using a mixed repeated measures analysis of variance (ANOVA) test: 2 (correct vs incorrect) x 2 (report vs no report) x 3 (rater experience level) on correctly assessed scans, reported as F and effect size partial eta squared η^2_p (Lakens 2013).

P values of ≤0.05 were interpreted as statistically significant.

3.4.3 Results

3.4.3.1 Test dataset characteristics

For the subjects included as test data, each group's mean age and standard deviation (SD) in years (y), and gender ratio was:

a) Left HS: 39.2y (13.5y) M:F 3:3;
b) Right HS: 44.7y (16y) M:F 4:5;
c) Bilateral HS: 42.3y (17.3y) M:F 2:3;
d) MR negative: 33.8y (10.1y) M:F 13:10.

There was no significant age difference between HS and MR-negative subjects (ANOVA $F(1,8)=1.83$, $p=0.159$).

Mean volume and qT2 values generated by the report pipeline for these test subjects are shown in Table 3-1, where left and right HS are combined as 'unilateral HS', volume ratio is calculated as unaffected : affected hippocampus and qT2 ratio as affected : unaffected hippocampus. The reference ranges for volume and qT2 ratios that were derived from the normative population are included for comparison.

Table 3-1. Volume and qT2 ratios for the test dataset, presented as unilateral HS, bilateral HS, and MR negative groups, with the normative reference ranges quoted for comparison.

Patient group	*Volume ratio % (range)* *Normative reference 88.9-110.6*	*qT2 ratio % (range)* *Normative reference 93.7-104.2*
Unilateral HS	72.8 (54.2 – 89.5)	107.8 (100.3 – 112.4)
Bilateral HS	86.4 (77.3 – 98.0)	99.2 (92.6 – 103.8)
MR negative	97.5 (85.9 – 110.1)	98.1 (94.9 – 102.2)

3.4.3.2 Detection accuracy

Detection accuracy (defined in section 3.4.2.6) was 87.5% without the quantitative report available. There was trend-level improved accuracy when raters had the report available, at 92.5% (p=0.07) with a moderate effect size (d=0.69) Table 3-2. The largest magnitude improvement effects were seen in the consultant radiologist and image analyst groups, with large effect sizes.

3.4.3.3 Lateralisation accuracy

Accurate lateralisation of pathology improved with the quantitative report. Of cases that raters correctly assessed to be abnormal, they incorrectly assigned the pathology (i.e., incorrectly chose right, left or bilateral HS) in 8.3% of cases without the report and in only 3.3% of cases when the report was available. Overall, lateralisation accuracy of HS by rater showed a trend-level improvement with the quantitative report, from 83.5% to 91.5%, p=0.075, with a moderate effect size (d=0.68).

3.4.3.4 Bilateral HS accuracy

When lateralisation of bilateral vs unilateral cases was compared, there was a significant improvement in overall accuracy in detection of bilateral cases with the report, p=0.028. There was significantly increased accuracy for detection of bilateral HS when using the quantitative report, from mean (SD) 74.4% (28.77) to 91.1% (17.64), p=0.042, d=0.70.

Table 3-2. Correct detection as normal or abnormal, combined for all raters and by rater group.

	Rater group	*Without Report Mean (SD)*	*With Report Mean (SD)*	*p-value*	*effect size, d*
Correct designation (normal / abnormal)	*Combined*	*87.3% (4.0)*	*92.5% (2.2)*	*0.06*	*0.73*
	1	92.2% (3.6)	96.1% (2.7)	0.30	1.23
	2	85.3% (6.7)	87.6% (3.6)	0.48	0.43
	3	84.5% (15.5)	93.8% (4.8)	0.27	0.81
Sensitivity	*Combined*	*87.5% (13)*	*90.0% (9.4)*	*0.25*	*0.41*
	1	96.7% (5.8)	98.3% (2.9)	0.74	0.37
	2	76.1% (16.7)	80% (8.7)	0.50	0.30
	3	90% (5)	91.7% (2.9)	0.42	0.41
Specificity	*Combined*	*87.4% (15)*	*95.0% (5.7)*	*0.14*	*0.54*
	1	88.4% (2.5)	94.2% (6.6)	0.18	1.15
	2	94.2% (5)	94.2% (5)	1	0
	3	79.7% (28)	95.7% (7.5)	0.31	0.78
Accuracy	*Combined*	*87.5% (9.0)*	*92.5% (5.0)*	*0.07*	*0.69*
	1	92.2% (3.6)	96.1% (2.7)	0.30	1.23
	2	85.9% (6.2)	87.6% (3.5)	0.60	0.33
	3	84.5% (15)	93.8% (4.8)	0.27	0.81

Rater groups: 1=experts; 2=trainees, 3=image analysts.

3.4.3.5 Rater agreement

3.4.3.5.1 Agreement between individual raters and the gold standard

Without the quantitative report, agreement between raters and the gold standard was moderate, Kappa (SD) = 0.74 (0.19). Agreement became strong when raters had the report available, Kappa (SD) = 0.86 (0.09), p=0.06, with a large effect size, d=0.81 (Table 3-3).

3.4.3.5.2 Agreement between raters

Reliability of assessment across raters showed some overall improvement in Cronbach's alpha from 0.452 without the report to 0.598 with the report

available. Agreement between raters also showed some improvement with the report, with the ICC for single measures increasing from 0.073 to 0.138 and for average measures from 0.417 to 0.591.

Table 3-3. Kappa scores for agreement of each rater and each group of raters with the gold standard.

Rater group	*Rater#*	*No QReport*	*With QReport*	*Net change*	*p value*	*Effect size, d*
Experts	1a	0.86	0.82	-0.04		
	1b	0.93	0.96	0.03		
	1c	0.78	0.96	0.18		
	Combined	*0.86*	*0.91*	*0.05*	*0.45*	*0.78*
Trainees	2a	0.86	0.82	-0.04		
	2b	0.69	0.80	0.11		
	2c	0.66	0.74	0.08		
	Combined	*0.74*	*0.79*	*0.05*	*0.38*	*0.68*
Analysts	3a	0.74	0.93	0.19		
	3b	0.30	0.78	0.48		
	3c	0.93	0.96	0.03		
	Combined	*0.66*	*0.89*	*0.23*	*0.22*	*1.13*
Total Mean (SD)		*0.74 (0.19)*	*0.86 (0.09)*	*0.12*	*0.06*	*0.81*

3.4.3.6 Rater confidence

Difference in confidence levels reported by raters for their assessments with report vs no report available, assessed in a series of paired t-tests (Table 3-4), showed that raters were significantly more confident in correctly assessing both normal ($p<0.01$, Hedges' $g_z=1.78$) and abnormal ($p<0.01$, $g_z=1.28$) scans when they had the quantitative report available.

A mixed ANOVA was used to assess the effect of the quantitative report on rater confidence accounting for rater experience level and whether a scan was normal or abnormal, in correctly assessed scans. A very large main effect of the quantitative report was found, showing that raters were more confident in their correct assessments with the report available [$F(1,6) = 102.65$, $p < .001$, effect size partial eta squared $\eta^2_p = .945$]. The effect of the quantitative report

on rater confidence was moderated by rater experience level, [QReport*Experience Interaction F(2,6) = 7.748, p =.022, η^2_p =.721], with a greater increase in confidence seen in the image analyst rater group [F(1,6) = 81.491, p < .001, η^2_p =.931]. Additionally, raters were more confident when correctly assessing abnormal scans than normal scans, irrespective of whether a quantitative report was available [F(1,6)=8.911, p =.024, η^2_p =.598].

*Table 3-4. Rater confidence when classifying normal and abnormal scans. * denotes statistical significance at <0.05.*

Confidence rating	*Δ (Report -No Report)*	*SD*	*95% Confidence Interval*	*t*	*df*	*p-value*	*effect size, g_z*
Overall Confidence	0.35	0.18	0.21-0.48	5.82	8	<0.01*	1.76
Normal	0.35	0.18	0.21-0.48	5.90	8	<0.01*	1.78
Abnormal	0.37	0.26	0.17-0.58	4.23	8	<0.01*	1.28
Correct normal	0.35	0.15	0.24-0.47	6.99	8	<0.01*	2.12
Correct abnormal	0.32	0.29	0.10-0.54	3.33	8	0.01*	1.00
Incorrect normal	0.14	0.37	-0.24-0.53	0.96	5	0.38	0.33
Incorrect abnormal	-0.31	0.24	-0.69-0.07	-2.61	3	0.08	-0.95

Δ = change in confidence level on 5-point scale. df: degrees of freedom. 'Correct normal' refers to raters' confidence when correctly assessing scans as normal.

3.4.4 Discussion

This clinical accuracy validation study was performed to test the effect of a quantitative MR biomarker report for detection of HS on the accuracy and confidence of image interpreters with differing levels of prior experience. Technically validated algorithms for quantification of hippocampal volume and qT2 were used to develop an automated report for single-subject biomarker reporting with accompanying normative reference data.

In summary, availability of the quantitative report was found overall to increase both assessment accuracy and rater confidence in detecting HS cases. This

was associated with strong effect sizes, despite sometimes reaching only trend-level significance, likely secondary to the small number of raters per group. Overall assessment accuracy and rater agreement with the gold standard both saw a large-effect improvement with report availability. All rater groups saw an increase in accuracy, and availability of reports increased accuracy of pathology lateralisation. Importantly, there was a significant increase in detection accuracy for bilateral HS cases.

Well-placed confidence in correct assessments increased significantly with use of reports. Rater experience level was an important modifier of rater confidence, with the largest report effect seen in the image analyst group.

There were very few instances of a rater making a correct assessment without the report and an incorrect one when they had the report available. This occurred in 1.7 cases per rater overall, and only in 1.3 cases per rater in the consultant radiologist group. These rates were less than the improvement shown per rater group with the report available.

The test dataset used covered a broad spectrum of HS disease severity, as the range of volumetry and qT2 data in Table 3-1 demonstrate. An important strength was the inclusion of subtle unilateral HS cases with atrophy ratios higher than 0.7, a threshold at which it has previously been shown that unaided visual detection can be very challenging (Reutens et al. 1996).

Accurate detection of MR features of HS is a central component to the management of patients with temporal lobe epilepsy. When these features are subtle or bilateral, accurate detection can be challenging. Previous validation studies for use of HS imaging biomarkers have shown increased assessment accuracy either alongside visual assessment or to outperform it when used in isolation. However these studies have not been designed to emulate the clinical environment for intended use, instead using arbitrary abnormality thresholds (Hu et al. 2018) or comparing quantification alone to visual assessment alone (Farid et al. 2012; Louis et al. 2020; Mettenburg et al. 2019). In contrast, this study considers the translational impact of combining standard visual assessment with quantification using clear tabulations, reference ranges and accessible graphical representations to assist the image interpreter.

Demonstrating that the combination of the two assessment processes improves HS detection and rater confidence supports it as a viable translational solution that could be incorporated as an adjunct into the clinical workflow.

The inclusion of distinct groups of raters with different levels of prior experience was useful in order to reflect the clinical situation and clarify which users would gain the most benefit from use of a quantitative tool. The largest improvements associated with the report were seen in the image analyst group, both for assessment accuracy and rater confidence. This was in keeping with the hypothesis that the least experienced raters would benefit the most from being able to reference the individual subject's results against a data from a reference normative population.

Interestingly, the expert group of raters saw increased agreement with the gold standard (Cohen's Kappa) with large effect sizes. This group has high baseline Kappa scores without the quantitative report, which may reasonably be explained by their years of previous experience and familiarity with HS compared to other rater groups. However their scores further increased with the quantitative report, suggesting a particular value to this group of raters in further assisting them in assessment of subtle or visually equivocal cases. These results suggest that the quantitative report may assist in levelling out the baseline discrepancy in rater expertise, affording individual patients with increased objectivity in the assessment received across imaging interpreters.

Whilst benefits were seen in the trainee radiologist rater group, they were less strong than those seen in the image analyst group. This may reflect a stronger reliance by image analysts on the report content, while trainees may be less inclined to assimilate the report information into their visual assessment in subtle cases. The improvements seen in the group of consultant raters seems to suggest that they were more able to integrate the report information into their visual assessment where necessary for improved detection of subtle cases than their more junior colleagues.

It is very interesting to see that use of the quantitative report in this study led to significantly increased assessment accuracy for bilateral HS cases. Bilateral

HS presents the dual challenge of being difficult to detect visually, due to the perceived preservation of symmetry and relative subtlety until there has been significant volume loss and/or sclerosis, and of being extremely important to detect accurately in order to avoid non-curative unilateral surgical attempts. Some surgical failures may indeed be due to subtle undetected bilateral disease that was not detected on imaging (Hennessy et al. 2000).

The use of a novel analysis and presentation method that quantifies volume and relaxometry values along the longitudinal axis of the hippocampus, in addition to global hippocampal values, presents an opportunity to more accurately localise a focal area of pathology, the identification of which is associated with a more favourable surgical outcome (Duncan and Sagar 1987). This would have to be investigated using histologically confirmed focal cases and comparing multidisciplinary decision making with and without the use of the quantitative report.

3.4.4.1 Limitations

Potential limitations to this study primarily related to the number of raters recruited to participate in the rating exercise and the number of cases that they were asked to assess. While overall accuracy saw a trend-level improvement with the use of the report, and large effect sizes were seen for most outcomes, it is likely that statistical significance was not reached due to the study being under-powered. When planning this study there were no examples of similar previous work available on which to base power calculations. Since the raters that were recruited achieved a high level of baseline accuracy without the report, a larger test subject set may also have been beneficial to show a significant benefit of the report. In addition, raters were performing at a high baseline without the report available, which may in part have been due to case selection, so it was more difficult to show a significant improvement effect.

Raters were not informed prior to the exercise how many HS cases they should expect to encounter, however it is possible that they were primed to expect them at a higher rate than normal clinical practice, due to the nature of the exercise. Additionally, it was not possible to fully recreate the expected clinical environment, as raters were blinded to any clinical referral information that they

might expect to have access to when assessing a patient's imaging. This was a necessary aspect of the study design, in order to attempt to isolate the effect of the report from any other informative data that may influence a rater's assessment. However in reality the reporter may integrate other information.

Case selection presented potential limitations. It was necessary to construct a test dataset with a robust clinical and/or pathological gold standard to allow for statistical analysis, however this may mean that by default the selected cases were those that were more clinically certain than cases that were subtle and did not reach surgery. This is difficult to avoid when a gold standard is necessary for study design. It was also possible that confirmed unilateral HS cases may have had unconfirmed contralateral pathology. It was also necessary to use expert neuroradiologist interpretation alongside additional clinical data to select bilateral HS cases. Quantitative reports were checked to confirm that the bilateral cases were captured adequately by the quantitative data; this assumes the diagnostic value of the quantitative report prior to its clinical validation.

Additionally, the non-HS subjects that were used consisted of MR-negative patients with epilepsy, who did not have established aetiologies identified, it is possible that there may have been undetected hippocampal pathology in this group. The test dataset had a wide age range but it was skewed towards younger adults, which is in keeping with the natural history of clinical presentation with HS.

All data used, including test and reference populations, was collected on a single scanner using a uniform imaging protocol. This is not reflective of the clinical reality of multiple scanners and protocols that may be routine in a clinical radiology department. While the results of this study can be taken to inform practice at this centre, application to more heterogeneous data and validation of multicentre reference ranges would be necessary to allow for more widespread adoption of the quantification pipeline. While multicentre reference values may lead to the tool becoming more representative, increased noise may be introduced. However, good interchangeability

between reference datasets has been shown in volumetric studies (Vinke et al. 2019).

Since the quantitative report presented both volumetric and qT2 measurements to the rater, this study did not address which of the two metrics is more useful and confers greater increase in accuracy and/or confidence. It would be interesting to consider whether a report presenting qT2 measurements only would achieve similar results to a combined report.

3.4.5 Conclusion

This clinical validation study represents a key proof-of-concept within the framework of development and translation of quantitative imaging biomarkers for clinical radiology use. It has shown that the use of single-subject quantification contextualised by normative reference data in an accessible graphical report format can improve accuracy of HS assessment, inter-rater agreement, and well-placed decision confidence.

Future directions for development of an HS biomarker tool may include a study assessing the effect of the report on the detection of focal pathology highlighted by the graphical representations of the hippocampal long axis. Additionally, as T2 relaxometry is not a conventional measurement outside of specialist epilepsy centres, a modified report could replace the T2 quantification used here with measurement of FLAIR signal instead, which would potentially make the tool more applicable to non-specialist centres. However, quantitative T2 measurements, which are not platform-specific, are likely to be more translatable between different centres than FLAIR quantification, which relies on local acquisition and calibration factors.

To build upon the results of this clinical validation study and follow the framework for clinical imaging biomarker validation that had been described, supervised integration into the local clinical service and in-use evaluation will allow the quantitative report to be applied to live clinical cases and be tested by neuroradiologists in their daily reporting workflow.

3.5 Deployment and in-service evaluation

Based on the results of the clinical validation study, and in line with the six-step QIB translational framework, the quantitative HS report has been systematically phased into early clinical deployment at a local level within the clinical neuroradiology service at the National Hospital for Neurology and Neurosurgery, Queen Square, London. This has been done by adopting a cautious and methodical roll-out strategy.

3.5.1 Pre-deployment modifications to the HS report

Subsequent to the clinical validation study, some aesthetic modifications were made to the HS report which had been identified as areas which could benefit from more clarity. In the table of global hippocampal volume and qT2 values, any abnormal values that were outside of the normative reference range were made to appear in bold text so that they could more easily be picked out. The table was also made larger by moving the QC information onto a second page of the report. Additionally, the hippocampal profile plots were ‘flipped’ from appearing as P-A on the original reports to now appearing as A-P, which is more naturally in line with radiological expectations. An example of the updated report is shown in Figure 3-3.

Instructions for use were added to the second page of the HS report, that were not included at the time of the clinical validation study. These instructions are important for clinical deployment as they provide the clinician with a context for how to interpret the report, important caveats and how to use it as a part of their radiological assessment.

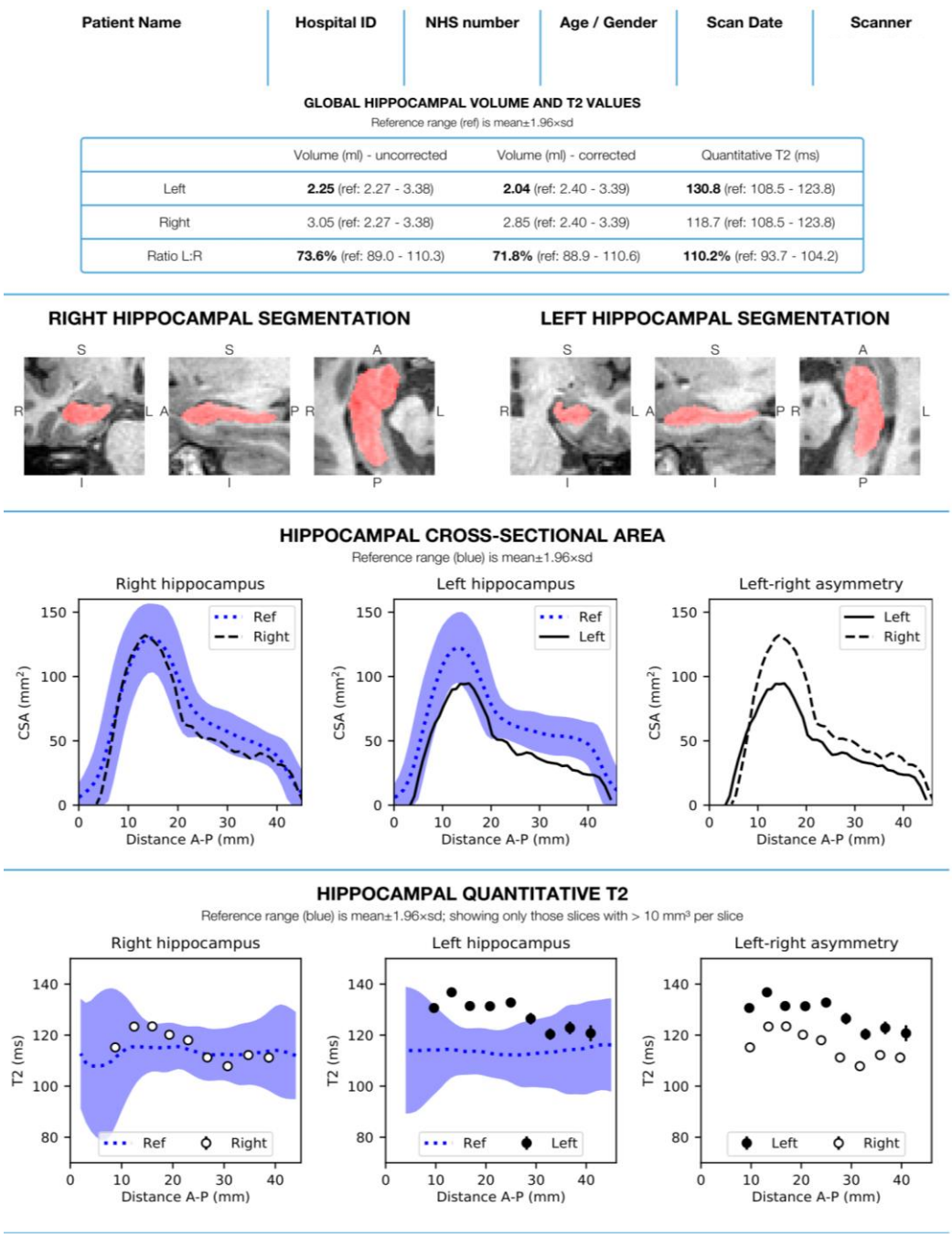


Figure 3-3. The updated HS report.

They also make clear that local clinical deployment of the tool is under the remit of an in-house exemption to the current medical device regulations, in lieu of official CE marking. The instructions for use were written with these aims in mind:

"This automated report has been developed as a tool to assist in interpretation of MRI examinations when a diagnosis of mesial temporal lobe epilepsy is suspected. An automated hippocampal segmentation technique (HippoSeg) has been used to generate hippocampal volumes and extract quantitative T2 measurements, and used as input to generate profiles of cross-sectional area and T2 along the hippocampal long axis (anterior-posterior). The quantitative report displays these values in the context of a normative range from healthy control subjects.

The report should be used as a supporting element of the radiologist's wider assessment, and does not replace conventional assessment or consideration of available clinical information and investigations. It only represents volumetric and T2 information, in the context of a normative population of 111 healthy controls (aged 17-67 years), and excludes any assessment of other pathology. It does not consider any historical imaging or other scans from this session, and should not be used to make assessment of changes in brain volumes between timepoints.

This report should only be used by radiologists trained in QNI Epilepsy Hippocampal Report interpretation. Measurements should not be compared to those provided by other software packages as algorithm compatibility cannot be assumed.

This tool is not CE-marked for clinical use and its use is allowed under the in-house exemption of the current regulations regarding medical devices.

For further detailed information about QNI report design and interpretation, please visit qni.cs.ucl.ac.uk."

3.5.2 Pre-deployment planning

3.5.2.1 Testing the clinical workflow

Prior to the incorporation of the HS report into the clinical workflow, a pilot test was undertaken using a shadow PACS system. This allowed for the testing of the analysis pathway, which allows an imaging study to be identified from the PACS system for processing with the HS quantification pipeline and a populated report to be sent back to PACS, to be viewed by the radiologist as an additional series in the patient's examination (Moggridge et al. 2020).

Ten real-time clinical cases were tested with this 'offline' deployment workflow and the resulting HS reports were reviewed by three consultant neuroradiologists alongside the patients' standard imaging series. Comments on the report output were made independently by the three consultants, followed by a group meeting where all the cases were reviewed and impressions were discussed.

All three reviewers were in agreement that the HS reports generated were processed correctly and returned credible quantitative information. Based on this consensus, a monitored roll-out of the HS report was authorised to begin with a group of pre-specified named consultant neuroradiologists within the department (see section 3.5.3).

3.5.2.2 Departmental training and awareness

A key element of planning for successful deployment, leading to effective uptake and staff engagement, was to ensure that staff from across the department were well informed of the background to the HS tool and the aims for its implementation. This was done by engaging the lead personnel for clinical workflow management, for example the lead radiographer for the department, in the deployment plans, as well as holding a dedicated presentation to which all staff members were invited. The presentation covered all stages in the technical development and clinical validation of the HS report, as well as the clinical deployment strategy. This enabled knowledge of the tool to be disseminated widely across the clinical department and allowed for direct engagement with questions and comments from end-users.

3.5.3 In-use evaluation

Ongoing in-use evaluation has been systematically performed from the outset of the limited departmental deployment. Clinical teams have been encouraged to request a quantitative report for patients with epilepsy who they are referring for MRI where a possible cause may be HS. These cases are filtered onto a separate reporting list which is assigned to four specific consultant neuroradiologists in the department. Every fortnight a review meeting is held between this consultant group to review the reports that have been processed and whether there are any issues identified. These meetings are also attended by the department's clinical scientists who process the requested reports so that any potential issues raised can be clearly communicated and promptly investigated.

To date, twelve quantitative reports have been issued in this clinical setting. Seven reports displayed normal hippocampal volumes and qT2 measurements. Three reports demonstrated RHS and two demonstrated LHS. There were no reports of bilateral HS. In all cases, the quantitative reports have aligned with the visual impressions of the reviewing consultant neuroradiologists. There have been no concerns raised by the radiologists or clinical scientists regarding the quantitative reports themselves. An issue of discussion has been how to increase clinical referrer engagement in order to increase the throughput of quantitative reports generated. This has been fed back to the neurology clinical lead for epilepsy to disseminate to referring colleagues.

Once a substantial case load has been reviewed by the pilot in-use team, and it is clear that there are no issues of concern, the workflow will be extended to all consultant neuroradiologists in the department for use in their routine reporting. In parallel, trainee radiologists spending time in the department and who have a high turnover rate will receive regular education and training on use and interpretation of the quantitative reports. Once its use is somewhat established across the department, a formal audit of use and impact of quantitative HS reports will be performed. These in-use evaluation steps are important precursors to larger-scale health socioeconomic impact

assessments that would ultimately be required to confirm the report's utility in a clinical setting.

4. Quantitative reporting for MRI in dementia

4.1 Introduction

The translational framework for imaging biomarkers that I set out in chapter 2 was followed by the example of its application to MRI in dementia. While I provided an outline in that chapter of the considerations for each step in the translational framework, this chapter will focus primarily on the development and clinical validation of a specific tool to be used by radiologists when assessing clinical scans for suspected dementia. Therefore, this chapter relates to steps 3 and 4 of the previously described QNI framework.

Structural MRI is the mainstay of general radiological practice for the investigation of patients with memory impairment and suspected dementia (Wattjes 2011). It can be used to delineate the differences in cerebral volume loss between normal aging and that which signals a superadded pathological neurodegenerative process, as well as to differentiate between specific dementia subtypes. The increased understanding of the biological basis of the different causes of dementia has led to diagnosis moving away from a diagnosis of exclusion towards one that can be described by clinical and imaging phenotypic patterns (Dubois et al. 2007).

Pathological atrophy of medial temporal lobe structures can be visualised on MRI and constitutes a core diagnostic feature of Alzheimer's Disease (AD) (Jack et al. 2002). Structural MRI can also contribute to differentiation between dementia subtypes (Vernooij and Smits 2012). Challenges exist where pathological changes are early and subtle, and where appearances overlap with changes that are related to non-pathological aging. Radiologists have designed visual rating scales as semi-quantitative imaging descriptors that can facilitate multidisciplinary assessments and patient follow-up, for example the medial temporal atrophy (MTA) scale (Scheltens et al. 1992). However, these scales are limited by their discrete nature and are not sensitive enough to describe early subtle changes (ten Kate et al. 2017; Pereira et al. 2014).

Quantification of a patient's imaging biomarkers and comparison to data taken from a healthy reference population could improve differentiation between normal and pathological MRI appearances (Brewer 2009), particularly as the focus shifts towards prophylactic and/or disease modifying interventions for

dementias in the future (McEvoy and Brewer 2010). Indeed, MRI volumetry is now well established in the research setting, with volumetric measurements often used as surrogate endpoints in clinical trials of potential therapeutic interventions for Alzheimer's Disease (Schwarz et al. 2019; Vandenberghe et al. 2017).

Advances in automated segmentation techniques have facilitated the processing of large normative reference datasets for single-subject comparison (Brewer 2009; Fischl et al. 2002). The use of normative reference data has been greatly enhanced by the availability of large online multi-site data repositories, which employ standardised imaging protocols and include patient and healthy control populations, the primary example being the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jack et al. 2008).

As a result of these advances, there has been a gradual translation from research application of automated volumetric MRI analysis to individual patients in clinical settings compared to an age-matched normative reference population. This could assist radiologists in their interpretation of patterns and extent of brain atrophy, and address clinical issues including the limitations of visual rating scales and the inter-rater variability inherent in routine visual evaluation. Use of quantitative volumetry reports has already been shown to increase clinical assessment accuracy (Hedderich et al. 2020; Vernooij et al. 2018) and to contribute to earlier detection of atrophy (Ross et al. 2013, 2015). These promising developments have been accompanied by a recent surge in proprietary software tools provided by private companies, whose products have received authorisation from regulatory approving bodies such as the Food and Drug Administration (FDA) and Conformité Européenne (CE) for this purpose.

Routine use of volumetric software for dementia assessment is currently low in clinical centres across Europe, at 5.7% in a recent survey of 193 centres, compared to 81.3% who routinely use the MTA scale (Vernooij et al. 2019). Significant translational barriers to the implementation of volumetric tools may partly explain this, including lack of imaging standardisation in clinical settings, and unclear clinical validation outcomes of proprietary software. Radiologists

may also be wary of potential additional time investment needed to adapt their routine workflow and whether the results of these quantitative tools will be easy to interpret. Despite their FDA and/or CE marking, many proprietary software solutions lack published clinical validation data. Currently there is no clear consensus for how these tools should be implemented in routine clinical reporting, particularly in relation to their impact on clinical management (Vernooij et al. 2018b). The translational framework that I set out in Chapter 2 aims to address the evidence gap by outlining a best practice development pathway for clinical use of QIB software tools.

In this chapter, I will present a quantitative report for use in dementia MRI assessment and its clinical validation process.

4.2 Methods

4.2.1 Automated brain volume quantification

Automated processing of brain scans for volumetric quantification has replaced cumbersome manual segmentation methods, and the developments of freely-available software over recent decades has facilitated this (Brewer 2009; Fischl et al. 2002). To construct a quantitative report that represents regional and global brain volumes, automated parcellation and segmentation methods were used.

4.2.2 Geodesic Information Flows

Brain segmentation was performed using the Geodesic Information Flows (GIF) algorithm (Cardoso et al. 2015). This technique is based on the premise of information extrapolation from a limited set of annotated data and applies it to a much larger sample. Specifically, it allows the propagation of categorical labels and probabilistic segmentations from one dataset to another. It provides fully automated multi-atlas segmentation and global and regional volumetry for T1-weighted MRI scans. It has been validated against manual segmentation methods applied to dementia and other neurodegenerative diseases (Bocchetta et al. 2016; Cardoso et al. 2015; Pardini et al. 2016).

Tissue segmentation for neuroimage analysis has classically been performed using probabilistic atlases to propagate and extrapolate information. However the observed intensities of an image are often insufficiently informative about the underlying tissue composition. This is due to inherent imaging limitations including inadequate signal to noise ratio (SNR), contrast to noise ratio (CNR), the presence of image artefacts and intensity non-uniformity (INU).

The addition of a priori spatial localisation information through coordinate mapping and propagation of anatomical priors can help to mitigate these issues. Anatomical priors can be generated by manual segmentation of a dataset and registration to a groupwise space, however this can obscure the morphological differences that represent natural variability and pathology.

Using several different sources via multi-atlas segmentation can be used to approach a good estimation of the actual tissue parcellation, however the

composition of the atlases is important, for example propagating information from young control data to an older population with Alzheimer's Disease would be problematic as the subjects in the source and target group are so morphologically dissimilar.

Several step-wise propagation algorithms have been devised, which allow for a low dimensional representation of the data to be used to propagate morphological similarity between datasets via intermediates, leading to increased segmentation accuracy (Wolz et al. 2010). GIF builds on this premise and provides a framework that propagates information between images using the geodesic path of a spatially-variant graph. The graph represents local patches and use of a restricted neighbourhood allows for increased accuracy and reduced bias. It is a general framework that can propagate different types of information including labels and image intensities.

The GIF algorithm used to generate these quantitative volumetric reports is based on a set of 35 subjects from the OASIS database that were manually segmented to produce 140 different labels. These labels contributed to eight different tissue classes: cortical grey matter (GM), supratentorial white matter (WM), cerebellar GM and WM, corticospinal fluid (CSF), deep GM, and pons.

4.2.3 Reference dataset

A reference dataset of subjects with no neurodegenerative disease was compiled and processed for contextualisation of individual subject brain volumetry.

It comprised volumetric T1-weighted scans of healthy controls from the Alzheimer's Disease Neuroimaging Initiative (ADNI), n=382, age range 56-90 years, augmented by the addition of younger subjects from the Track-HD study cohort, n=79, age range 30-65 years. In total, the normative dataset comprised 461 subjects, 51.4% female, with a mean age of 70.09 years (SD=12.05).

Acquisition parameters for each of the study cohorts were as follows.

ADNI-2 protocol:

sagittal plane acquisition using an MP-RAGE/IR-FSPGR pulse sequence on a range of 3.0T MRI systems with the following parameters: 8-channel coil, TR=400ms, TE=min full, flip-angle=11°, slice thickness=1.2 mm, resolution=256 × 256 mm and FOV=26 cm.

Track-HD protocol:

T1-weighted image volumes were acquired using a 3D MPRAGE acquisition sequence on 3.0T Siemens or a Phillips MRI system with the following imaging parameters: TR=2200ms (Siemens)/ 7.7ms (Philips), TE=2.2ms (S)/3.5ms (P), FOV=28cm (S)/ 24cm (P), matrix size 256x256(S)/224x224(P), 208(S)/164(P) sagittal slices to cover the entire brain with a slice thickness of 1.0 mm with no gap.

4.2.4 Quantitative report design

The quantitative report brings together the different elements of the processing pipeline and displays an individual subject's quantitative results in the context of the normative reference dataset (Figure 4-1). The report displays the subject's demographic information at the top, including identification number, age, date of scan and scanner type. Below this it displays snapshots of the hippocampal segmentations that have been performed using GIF, overlayed in red over the subject's T1-weighted MRI scan, so that the report user can perform their own brief visual assessment of the segmentation output. The snapshots are accompanied by the numerical percentile values for left and right hippocampus which are calculated in reference to the normative population.

Alongside this is a graph displaying the brain parenchymal fraction (BPF) for the individual subject, plotted on a graph displaying BPF by age for the reference population. The individual patient's BPF is displayed as a large red dot and each subject in the reference population is represented by a small black dot. The mean and standard deviation lines for the reference population are displayed on the graph and a traffic-light system of colours is used to show the areas of the graph that are within the healthy range (green), within one

standard deviation from the healthy range (yellow) and more than two standard deviations away from healthy (red).

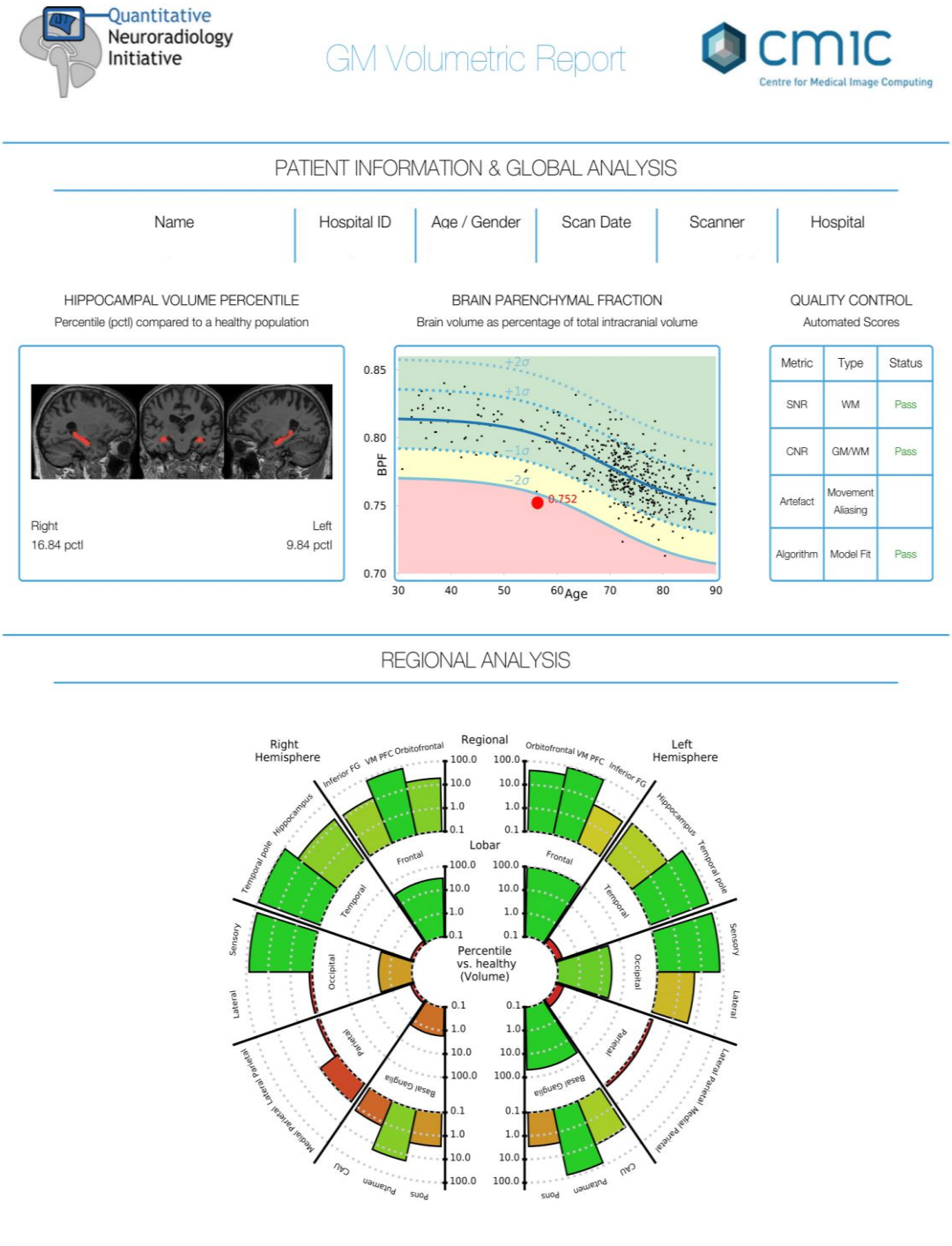


Figure 4-1. Quantitative report for a patient with AD.

Completing the report is a detailed representation of the volume percentiles of specific brain regions. This information is communicated in the form of a ‘bullseye’ plot, which is a means to display complex 3D data in a visually simplified format, as has been done elsewhere to represent lesion volume information (Sudre et al. 2018). These are expressed as percentile estimates against a Gaussian distribution approximation of the specific regional volume derived from the healthy reference population, controlling for age, gender and total intracranial volume (TIV). Percentiles were calculated based on the use of a variant of a generalised logistic function to predict the values from the healthy reference database as a continuous variable. This allowed the computation of the cumulative distribution function of an individual’s measured volumes in reference to the reference population. The inner ring of the bullseye represents volume percentiles at the lobar level and the outer ring expands on specific sub-regional areas within each of the lobes that are important to consider when assessing an MRI scan for the presence of atrophy.

4.3 Credibility study

4.3.1 Aims of the credibility study

The credibility study aims to assess the credibility of the described automated quantification pipeline and report as an aid to MRI interpretation in patients with dementia. This will be done by assessing the accuracy of whole brain and regional segmentations, evaluating the credibility of the information displayed in the quantitative report, and evaluating whether the report is reflective of the clinical impression made independently by expert radiologists.

4.3.2 Methods

We selected 10 subjects with a radiological diagnosis of either Alzheimer's Disease (AD) or frontotemporal dementia (FTD), and 10 healthy controls. T1-weighted MRI data was collected between 2014 and 2016 at the National Hospital for Neurology and Neurosurgery, Queen Square, London. Demographic information and pathology reports of CSF analysis for tau and amyloid beta were also collected.

Data was post-processed with the previously described segmentation software tool GIF (Cardoso et al. 2015). As described, GIF was used to calculate whole brain and regional volumes for each subject using a multi-atlas segmentation approach. Individual subject values were expressed as percentile estimates against a Gaussian distribution of previously obtained normative grey matter volumes of healthy control subjects. The normative data was derived from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Track-HD cohort.

Data was then presented in a report format, with graphs for whole brain GM volume plotted against normative volumes, and a 'bullseye' representation of GM volume percentiles by brain lobe and pertinent sub-regions.

An expert neuroradiologist with special interest in dementia assessed the T1-weighted images for each subject and made their clinical impression. The expert also evaluated the accuracy of the brain segmentations by viewing them as overlays on the T1 images using an image viewing platform niftiMIDAS (Clarkson et al. 2015). Finally, the expert appraised the QNI report for whether it was reflective of their clinical impressions, by rating whole brain and regional

GM volumes as credible or doubtful. The reports were also assessed for any technical or presentation issues.

4.3.3 Results

4.3.3.1 GM whole brain and regional volumes

The expert neuroradiologist assessed T1 studies from 20 subjects. Nineteen of the 20 were rated as credible for whole brain volume. Only 12 of the 20 subjects' regional GM volumes were rated as credible. It was noted that the range of normative data was heavily weighted to subjects over the age of 65, and this was thought to be causing the doubtful GM regional volumes for subjects who were in the younger age ranges. To rectify this, 100 additional control subjects below the age of 65 were added to the normative database, taken from the track-HD cohort, and the same 20 subjects' QNI reports were re-generated.

Following this, the same rating process was carried out using the new QNI reports. All 20 brain parenchymal fraction measurements appeared credible. Regional volume percentiles appeared fully credible for 14 of the twenty. Six cases had minor areas of discrepancy, including five which somewhat overestimated the amount of atrophy seen visually in certain regions and one which somewhat underestimated visual hippocampal atrophy.

All the cases where regional volumes appeared doubtful were reviewed with their regional normative data fits. The statistical 'best fit' for the normative data was trialled and a polynomial fit was superior (Figure 4-2). These cases were then assessed to be credible.

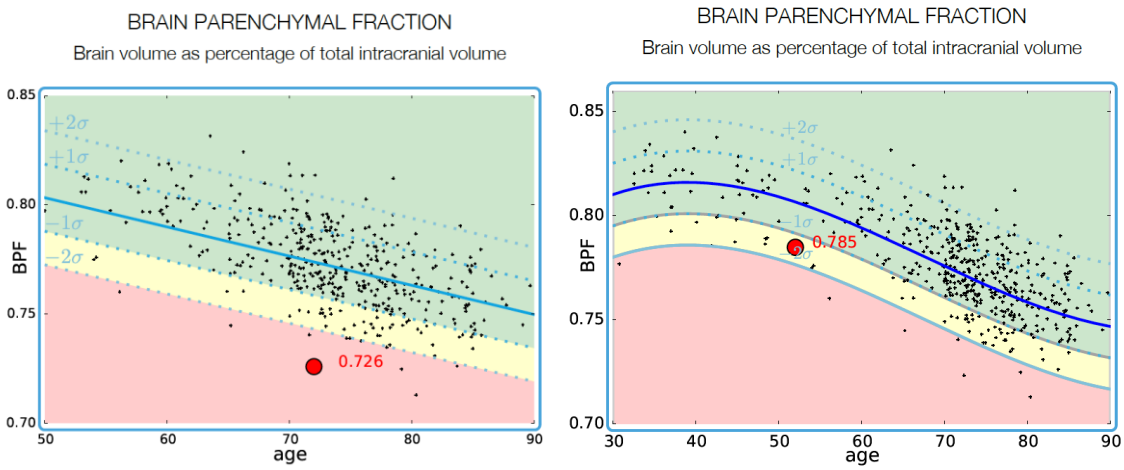


Figure 4-2. Brain Parenchymal Fraction (BPF) plots. Left – Original BPF graph displaying linear model fit with data in age range 50 – 90. Right – updated BPF graph displaying a polynomial fit with the addition of data in age range 30 – 90.

4.3.3.2 Technical or presentation issues

Reports were technically accurate in most cases. In one instance the whole brain GM volume for a subject was lower than our parameters allowed for and so was not visible on the graph. This issue was rectified when we added more cases to our normative database and the data was refitted.

4.3.4 Outcomes

Having carried out this credibility study, the quantitative pipeline was applied to a larger set of test cases to conduct a clinical validation study with a wider group of raters, which assessed the potential benefits to clinical accuracy and reporting confidence that may be seen with the use of a quantitative report.

4.4 Clinical accuracy validation study

4.4.1 Aims and hypothesis

This accuracy study was performed to assess the effect of the quantitative brain volume report on clinical accuracy of interpretation, across two diagnostic steps and three neuroradiological levels of experience. It was hypothesised that the use of the quantitative report would decrease interrater variability whilst increasing assessment specificity, sensitivity, accuracy, and confidence (a) for determining the presence of volume loss and (b) for determining the differential diagnosis of AD or FTD; and that the report's effect would be identifiable across the three levels of rater experience.

4.4.2 Methods

All subjects in the test dataset gave their informed consent for all parts of the study under research ethical approval by the UCL Queen Square Institute of Neurology ethics committee.

4.4.2.1 Test dataset

A test set of MRI scans was established from 45 subjects scanned at the NHNN, using three different 3-T MRI systems:

1. 41 scans on Trio MRI scanner, (TR=2,200 ms, TI=900 ms, TE=2.9 ms, acquisition matrix=256×256, FOV=26 cm, spatial resolution=1.1 mm),
2. 3 scans on Skyra MRI scanner (TR=2,200 ms, TI=900 ms, TE=2.9 ms, acquisition matrix=256×256, FOV=26 cm, spatial resolution=1.1 mm), and
3. 1 scan on Prisma MRI scanner (TR=2000 ms, TI=850 ms, TE=2.93 ms, acquisition matrix=256×256, FOV=26 cm, spatial resolution=1.1 mm).

Sixteen subjects had been diagnosed AD, with CSF levels of beta-amyloid 1-42 <550 pg/mL and tau:amyloid ratio >1, and fourteen with FTD, based on clinical evaluation and CSF markers. Fifteen subjects who had been referred to the specialist memory clinic with subjective memory concerns but who were deemed to fall within normal ranges upon neurological, cerebrospinal fluid

(CSF) and imaging assessment were used as a non-neurodegenerative sample.

4.4.2.2 Assessment exercise

The study involved three groups of raters each containing three people: consultant neuroradiologists; radiology trainees with an interest in neuroradiology; and non-clinical image analysts who had experience with dementia MRI, either as specialist radiographers or non-clinical dementia research fellows. Raters were invited from a range of centres both in the UK and across Europe, to ensure a broad representation of previous training and experience. The rating exercise was hosted on a website platform to allow for remote participation, and it was designed to replicate the standard working environment that a radiologist would expect when assessing scans. Raters were blinded to all clinical and demographic information relating to the study subjects apart from age and gender which was displayed alongside each case. The 45 subject scans were presented to the raters twice, in a randomly generated order that was unique to each rater, once with and once without the quantitative report available alongside the images. Therefore, the assessment consisted of ninety rating episodes in total. At each rating episode, the rater was asked to assess whether the scan was normal or abnormal, in terms of appearance of brain volume for age, and to give their confidence level for that assessment on a scale of 1 (very uncertain) to 5 (very confident). If the rater had assessed the scan as abnormal, they were then asked whether they assessed the abnormality as AD or FTD, and again asked for their confidence in this assessment on the 1-5 scale. Ratings were completed over a two month period, collected via the website platform and subsequently analysed.

4.4.2.3 Instructions to raters

These instructions were supplied to raters as an introduction page on the assessment website:

We have developed a tool to assist MRI interpretation for patients with suspected dementia. It is an automated grey matter (GM) segmentation technique which calculates whole brain and regional GM volumes from a

subject's MRI scan, and presents a QNI report in the context of a normative range of control subjects.

We are keen to test how helpful it is for radiologists and other imaging staff to have this quantitative information. In this exercise, you will see a mixture of MRI scans from individuals with Alzheimer's dementia (AD), frontotemporal dementia (FTD), and those who are cognitively normal. We have chosen to focus on AD and FTD at present due to their distinctive patterns of atrophy.

You will see each scan in three planes and you will be able to scroll through the images as you would normally. You will see the same scans twice, once with and once without a QNI report, in a randomly mixed order. Please use the QNI report where it is available to assist you in making your judgement, otherwise please rely on your visual interpretation as you would normally.

An example QNI report appears on the right. The graph at the top indicates the subject's whole-brain volume. The 'bullseye' plot below shows the percentile of the GM volume in each lobe and important sub-regions, either higher (green) or lower (red) percentile.

For each scan, we will ask you to give your assessment:

1. *Your overall impression – normal or abnormal*
2. *If you think the scan is abnormal, specify your diagnosis – AD or FTD*
3. *How confident you feel in your diagnosis on a scale 1-5, 1=not at all confident, 5=very confident.*

*Once you have finished looking at a scan, click **NEXT** and your assessment will be saved. To stop and come back to the exercise at any time, click on **MENU** in the top right hand corner of the screen. The next time you log in, the website will take you to the next scan to be reported. You can zoom in and out on the report using your mouse wheel.*

Thank you again for your participation, we look forward to sharing the results with you!

4.4.2.4 Statistical analysis

We explored the effects of availability of the quantitative report to raters on their accurate assessment of the presence of pathological atrophy (normal versus abnormal) and then the ability to differentiate between AD and FTD. Signal-detection indices were used to assess the following ratings:

(a) True positive for AD/FTD (correctly assessed as abnormal),
(b) True negative' for healthy controls (correctly assessed as normal)
(c) False positive' for healthy controls (incorrectly assessed as abnormal), and
(d) False negative for patients (incorrectly assessed as normal).

Using these metrics, diagnostic sensitivity, specificity, and accuracy were calculated and expressed as percentages:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + False\ Positives + False\ Negatives}\ x\ 100$$

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives}\ x\ 100$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives}\ x\ 100$$

Counts of correctly and incorrectly assessed scans with and without the quantitative report available were then analysed using the McNemar test. Mean diagnostic accuracy, specificity, and sensitivity across the two conditions (report present versus absent) were assessed with paired t-tests.

Cohen's kappa was calculated to assess agreement between raters' evaluations and the gold standard, while accounting for chance agreement. To further assess the effect of quantitative report availability on consistency and reliability among raters, Cronbach's alpha and intraclass correlation coefficients were also calculated.

Confidence ratings with the report versus without were calculated as a grand mean per rater and for each true disease type (normal, AD, FTD) and were assessed using paired t-tests.

In an exploratory analysis of ratings, the effect of the presence of the quantitative report on correct assessment and rater confidence was explored in terms of whether the rated scan was in fact normal or abnormal as well as how experienced the rater was. This was done using a four-way mixed ANOVA (report available × normality × correctness × experience level), to assess the interaction between these factors. All statistical analyses were performed with SPSS version 24.

4.4.3 Results

4.4.3.1 Test dataset characteristics

Subject groups were age matched (years, mean (SD); gender male:female): non-ND group 60 (8.7), 4:11; AD group 61.7 (6.6), 9:7; and FTD group 59.9 (7.3), 11:3. CSF confirmed that AD subjects had reduced Aβ 1-42 and raised mean Tau levels. Mean MMSE was significantly lower for AD ($p<0.001$) and FTD ($p=0.03$) when compared with the non-neurodegenerative (non-ND) group. Mean disease duration (time from first reported symptom to MRI) was (years mean, (SD)) 2.7 (1.6) for the AD group and 3.5 (2.4) for the FTD group. Test dataset characteristics are shown in Table 4-1.

Table 4-1. Test subject dataset characteristics.

	Non-ND (n=15)	AD (n=16)	FTD (n=14)	Total (n=45)
Age in y, mean (SD)	60 (8.7)	61.7 (6.6)	59.9 (7.3)	60.6 (7.4)
Gender male:female	4:11	9:7	11:3	24:21
Mean Aβ 1-42 (pg/ml)	878.8	393.3	747.7	-
Mean Tau (pg/ml)	373.8	855.2	302.6	-
MMSE, mean (SD)	26.9 (4)	20.5 (6.4)	22 (9.1)	-
Disease duration in y, mean (SD)	-	2.7 (1.6)	3.5 (2.4)	-

4.4.3.2 Assessment accuracy

4.4.3.2.1 Assessment of cases as normal or abnormal

Availability of the quantitative report significantly improved assessment sensitivity (p=0.015) but did not significantly change assessment specificity or accuracy. A beneficial moderate effect size was seen for assessment accuracy (d=0.53). Accuracy calculated by rater group showed a significant increased assessment accuracy with the quantitative report for the consultant neuroradiologist group, from 71% to 80% accuracy, p=0.02 (Table 4-2).

*Table 4-2. Sensitivity, specificity and accuracy for normal vs abnormal rating across all experience levels, both with and without the quantitative report. * denotes statistical significance at <0.05.*

Metric	*Experience Level*	*Without report Mean (SD)*	*With report Mean (SD)*	*p-value*	*d effect size*
Sensitivity	*Consultant*	68.9% (5)	80% (10)	0.13	1.4
	Registrar	75.5% (8.4)	81.1% (1.9)	0.3	0.8
	Image analyst	70% (25.1)	85.5% (10.1)	0.23	0.9
	All groups combined	71.5% (13.8)	82.2% (7.6)	0.015*	1.03
Specificity	*Consultant*	75.6% (3.8)	80% (13.3)	0.52	0.43
	Registrar	82.2% (10.1)	68.8% (25.2)	0.37	-0.6
	Image analyst	77.7% (3.8)	68.9% (13.8)	0.45	-0.52
	All groups combined	78.5% (6.4)	72.3% (16.8)	0.3	-0.37
Accuracy	*Consultant*	71.1% (2.2)	80% (2.2)	0.02*	4
	Registrar	77.7% (3.8)	77% (9.2)	0.87	-0.1
	Image analyst	72.6% (17.9)	80% (2.2)	0.5	0.46
	All groups combined	73.8% (9.5)	79% (5.1)	0.15	0.53

4.4.3.2.2 Assessment of cases as AD or FTD

Availability of the quantitative report significantly improved sensitivity for detecting AD cases across all raters (p=0.002) (Table 4-3). No significant improvements were seen for FTD detection (Table 4-4).

In absolute terms, correct assessments of AD and FTD cases increased with the quantitative report by 6.9% and 5.6% respectively, however these increases were not statistically significant.

*Table 4-3. Sensitivity, specificity and accuracy for AD vs normal rating across all experience levels, and percentage of correct assessments for AD, both with and without the quantitative report. * denotes statistical significance at <0.05.*

Metric	*Experience Level*	*Without report Mean (SD)*	*With report Mean (SD)*	*p-value*	*d effect size*
Sensitivity	*Consultant*	61.3% (12.9)	75.8% (17)	0.05	0.96
	Registrar	79.3% (12.7)	83.9% (4.7)	0.42	0.48
	Image analyst	61.7% (45.1)	71.1% (44.4)	0.01*	0.22
	All groups combined	67.4% (25.8)	76.9% (24.5)	0.002*	0.37
Specificity	*Consultant*	79.6% (8)	82% (14.6)	0.73	0.2
	Registrar	83.1% (7.5)	78.1% (7.8)	0.46	-0.65
	Image analyst	86.6% (4.2)	77.9% (12.9)	0.31	-0.9
	All groups combined	83.1% (6.6)	79.3% (10.7)	0.3	0.42
Accuracy	*Consultant*	70.7% (3.2)	79.2% (10.9)	0.07	1.05
	Registrar	80.3% (2.3)	78.9% (7.4)	0.76	-0.25
	Image analyst	75.8% (16.8)	77.9% (4.2)	0.84	0.17
	All groups combined	75.5% (9.6)	78.7% (4.3)	0.38	0.43
Correct % AD diagnoses		58.1% (3.4)	65% (4.1)	0.128	0.56

*Table 4-4. Sensitivity, specificity and accuracy for FTD vs normal rating across all experience levels, and percentage of correct assessments for FTD, both with and without the quantitative report. * denotes statistical significance at <0.05.*

Metric	*Experience Level*	*Without report Mean (SD)*	*With report Mean (SD)*	*p-value*	*d effect size*
Sensitivity	*Consultant*	57.3% (4.1)	57.2% (6.2)	0.93	-0.01
	Registrar	36.5% (7.8)	35.2% (23.8)	0.94	-0.07
	Image analyst	46.9% (24.9)	58.3% (20.2)	0.1	0.5
	All groups combined	46.9% (16)	50.3% (19.5)	0.52	0.19
	Consultant	89.2% (9.4)	95% (6.5)	0.19	0.71

Specificity	*Registrar*	91.1% (9.7)	77.7% (32.7)	0.42	-0.55
	Image analyst	75.5% (28.9)	85.5% (14.5)	0.46	0.43
	All groups combined	85.2% (17.6)	86.1% (19.7)	0.89	0.04
Accuracy	*Consultant*	73.6% (4.9)	75.9% (3.6)	0.09	0.53
	Registrar	70.5% (11)	65.2% (22.8)	0.52	0.29
	Image analyst	69.1% (15.9)	72.6% (14.3)	0.41	0.23
	All groups combined	71.1% (10.2)	71.2% (14.4)	0.95	0.01
Correct % FTD diagnoses		38.6% (2.2)	44.2% (2.7)	0.367	0.31

4.4.3.3 Assessment confidence

When rating cases as normal versus abnormal, a four-way mixed ANOVA (report available × normality × correctness × experience level) showed availability of the report conferred a significant increase in confidence when incorrectly rating abnormal scans ($p=0.02$, $F(1,8)=7.918$), with a small effect size ($\eta 2p= 0.497$). This effect did not vary across rater groups.

Raters were also significantly more confident with the report available than without, regardless of correctness ($p=0.03$, $F(1,8)=6.64$, $\eta 2p=0.453$). They were significantly more confident when correctly assessing a scan, regardless of report availability ($p=<0.01$, $F(1,8)=112.43$, $\eta 2p=0.934$) and also when they were rating abnormal versus normal scans, also regardless of report availability ($p=<0.01$, $F(1,8)=21.68$, $\eta 2p=0.73$).

4.4.3.4 Agreement between raters and the gold standard

For assessment of scans as normal versus abnormal, and for differentiating between AD and FTD, there was a significant increased agreement (Cohen's Kappa) for raters in the consultant group with the gold standard ($p=0.038$ and $p=0.04$ respectively), Table 4-5 and Table 4-6.

*Table 4-5. Kappa scores for normal/abnormal assessments across all experience levels, both with and without the quantitative report. * denotes statistical significance at <0.05.*

Experience Level	*Rater#*	*No Report*	*With report*	*Net change*	*p-value*
Consultant	A1	0.400	0.586	0.186	0.038*
	A2	0.469	0.571	0.102	
	A3	0.381	0.492	0.111	
Registrar	B1	0.455	0.211	-0.244	0.68
	B2	0.522	0.571	0.05	
	B3	0.613	0.667	0.054	
Image analyst	C1	0.169	0.531	0.362	0.66
	C2	0.746	0.556	-0.19	
	C3	0.492	0.557	0.065	
Overall Mean (SD)		0.48 (0.17)	0.52 (0.13)	0.04	0.34

*Table 4-6. Kappa scores for agreement between rated diagnosis and clinically/CSF-confirmed AD and FTD diagnoses across all experience levels, both with and without the quantitative report. * denotes statistical significance at <0.05.*

Experience Level	*Rater#*	*No Report*	*With report*	*Net change*	*p-value*
Consultant	A1	0.432	0.531	0.099	0.04*
	A2	0.45	0.498	0.048	
	A3	0.335	0.434	0.099	
Registrar	B1	0.381	0.22	-0.161	0.56
	B2	0.326	0.428	0.102	
	B3	0.494	0.391	-0.103	
Image analyst	C1	0.02	0.176	0.156	0.28
	C2	0.529	0.496	-0.033	
	C3	0.396	0.529	0.133	
Overall Mean (SD)		0.37 (0.15)	0.41 (0.13)	0.037	0.39

4.4.3.5 Agreement and reliability across raters

When assessing scans as normal versus abnormal, Cronbach's alpha for agreement across all raters showed improvement in overall rating reliability from 0.886 to 0.925 with the quantitative report available, which can be interpreted as an improvement from 'good' to 'excellent' agreement. The intraclass correlation co-efficient, assessed using mixed two-way ANOVA across raters, increased from 0.454 to 0.563 for single measures and from 0.882 to 0.921 for average measures with the quantitative report available.

4.4.4 Discussion

This clinical accuracy study was designed to assess the impact of an automated quantitative volumetric report on user assessment and confidence. An established segmentation algorithm, GIF, was used to develop a quantification pipeline which produces a summary report that brings together patient demographic information, hippocampal segmentation, brain parenchymal fraction, and global- and region-specific brain volumetry contextualised against a normative population.

This study assessed the effects of this novel quantitative volumetric report on assessment sensitivity, specificity, accuracy and confidence across three groups of raters with different levels of previous image interpretation experience. Availability of the quantitative report while interpreting MRI scans increased the sensitivity of detecting volume loss across all raters and improved both the accuracy and agreement with the gold standard in the consultant group of raters. The report also improved sensitivity for detecting AD for the image analyst group and for all raters combined, an effect that was not seen for FTD discrimination.

Variability in accuracy, sensitivity, and kappa scores for detecting volume loss all reduced with the report. In absolute terms, classification accuracy increased overall by over 5%. Given the documented increases in dementia prevalence in recent years and its future projections (Ahmadi-Abhari et al. 2017), this figure could be of clinical importance if confirmed in a larger study population.

Significant improvements seen in the assessments made by the neuroradiologist group of raters suggests that experienced image readers are well placed to assimilate and make use of the information provided in the quantitative report. Conversely, it is possible that less experienced neuroradiologists and non-clinical image analysts were over-reliant on the report for determining abnormality, as suggested by an overall decrease in specificity, although this was not statistically significant.

When assessing an MRI study for the presence of a neurodegenerative disease, neuroradiologists make an assessment of possible disease features to determine the presence of abnormality. In this study we have suggested that reporters may first assess for the presence of abnormal volume loss. They may then concentrate on its distribution and interpret the specific pattern of pathological atrophy to be indicative of a certain disease type, such as AD or FTD. The results from this study may be interpreted in this two-stage analysis context, which in real-world settings are not independent from each other and are performed as a single analysis. Providing the quantitative report increased the sensitivity of the first stage (i.e., the assessment of volume loss across all raters) and improved the accuracy and agreement among the consultant group. For differentiating the pattern of atrophy present, the report improved sensitivity for AD in the image analysts and for all raters combined but not for FTD. The provision of a quantitative report that provides objective measures to reproducibly assess volume loss and leading to increased rater agreement could be clinically useful in terms of screening, diagnosis and training across a clinical dementia service.

The limited effects on the differential diagnosis on FTD may be explained by the low mean age of patients and relatively short disease durations in the study population. Cases were possibly more subtle or early in the disease course, before specific patterns of atrophy were prominent, which may interact with raters being relatively less familiar with MRI appearances of FTD than those of AD. However, it is also important to identify atrophy in younger patients while it is still subtle, and it is in these cases especially where a quantitative report may help to reduce subjective visual disagreement.

Interestingly, rater confidence in detecting volume loss and differentiating AD and FTD was not significantly affected by availability of the quantitative report. Significantly increased confidence was unexpectedly shown when incorrectly diagnosing volume loss (i.e., the report introduced false confidence) independent of experience level. One potential explanation is that raters based their incorrect diagnosis on visual inspection alone and used the report to reinforce their diagnosis.

More work needs to be done to understand and mitigate such findings. Rigorous validation is needed before clinical adoption. The importance of appropriate training to avoid over-reliance on diagnostic aids, as well as the careful planning and monitoring of how tools such as this quantitative report are used when implemented in clinical practice, are key to successful translational development. Rather than being viewed as a new gold standard containing all the answers, quantitative reports should be framed as support tools which cannot replace neuroradiological experience, and users should be wary of over-reliance.

As discussed in Chapter 1, several proprietary quantitative tools exist for the assessment of dementia, for example Neuroquant (cortechs.ai 2021) and icobrain-dm (icometrix 2021). Technical validation of their segmentation algorithms has been performed in comparison to other segmentation procedures, with promising results (Brewer 2009; Struyfs et al. 2020). However, systematic assessments of their impact on clinical image interpretation by neuroradiologists have not been published, despite both tools being FDA and CE approved.

There is a noticeable lack of clinical validation studies in the literature for volumetric neuroradiological tools. Those that have been conducted have typically used only two raters, and do not always assess performance of visual inspection combined with quantitative report use. A recent study showed use of non-proprietary quantitative reports improved the identification of patients versus healthy controls for one of two raters, while both raters improved in the differential diagnosis in a group of patients with AD and FTD dementia (Hedderich et al. 2020). In another study it was shown that combining

quantification of lobar and hippocampal volumes with visual inspection improved the diagnostic accuracy of two experienced neuroradiologists, but that providing raters with the quantification results alone reduced their interpretation accuracy (Vernooij et al. 2018). This finding reinforces that quantitative reports should be used as adjuncts that complement visual assessment, and that clinical validation of these tools should aim to reflect the clinical environment as closely as possible.

4.4.4.1 Limitations

This study was somewhat limited in statistical power, potentially due to the subject sample size or the number of raters used. The sample size of 45 subjects was in line with other similar studies using between 36 and 52 subjects (Heckemann et al. 2008; Hedderich et al. 2018; Vernooij et al. 2018). The use of nine raters within three experience levels enabled the assessment of the effect of prior reporting experience when introducing quantitative reports. Similar work has used only 2 raters (Heckemann et al. 2008; Vernooij et al. 2018) or a maximum of 3 raters (Hedderich et al. 2018).

The performance of the non-clinical image analyst group was unexpectedly heterogeneous, likely due to disparity in previous experience level. The variability in the results within the image analysts and registrar groups could also reflect an over-reliance on the report by less experienced reporters. This study therefore underlines the importance of considering sample sizes and rater groups when validating such quantitative diagnostic aids. Power calculations show that future work will need to involve a larger number of raters to better assess the effects of the report on assessment performance, and the moderators of this.

Control subjects were recruited from a clinical population who all presented with subjective neurological complaints. It is possible that this group contained subjects with subtle undetected pathologies, which may have affected rater performance. This was, however, a conscious choice to reflect the clinical scenario that reporters would face in memory clinic services. Finally, the incidence ratio (controls:AD:FTD), forced-choice nature, and lack of background clinical data in this study do not adequately reflect routine

neuroradiological assessment. In particular it may have been useful to remove the forced-choice nature of the assessment by providing a third diagnostic category of ‘abnormal, non-AD non-FTD’.

4.4.4.2 Conclusions

This clinical accuracy study demonstrates that quantitative volumetric reports providing single-subject results referenced to normative data can improve the sensitivity, accuracy, and inter-observer agreement for detecting pathological volume loss and AD. The largest beneficial effect of the quantitative report was seen in the consultant group, suggesting that this group was best placed to assimilate and make use of the information provided by the reports. The differing effects between three rater experience levels highlight the need for future work to clarify the potential benefits and limitations of these reports, and the importance of rigorous validation before clinical adoption. Statistical power was low, but the effect sizes seen across accuracy and sensitivity were moderate-to-large in favour of a beneficial report effect. Importantly, reduced variability in sensitivity, accuracy, and rater agreement scores was also noted. This study can inform power calculations and study design for future research in this field.

4.5 Lessons and future work

4.5.1 Report changes

Due to the mixed results of the clinical accuracy study, some changes were made to the quantitative report presentation. The changes were focused on simplifying the regional volume ‘bullseye’ plot, so that it featured fewer numbers and used a wider spectrum of colours, to highlight a wider range of volumetric information. The new report also features a scale underneath the bullseye plot so that users can reference the colour system against the numerical scale. The updated report is shown in Figure 4-3.

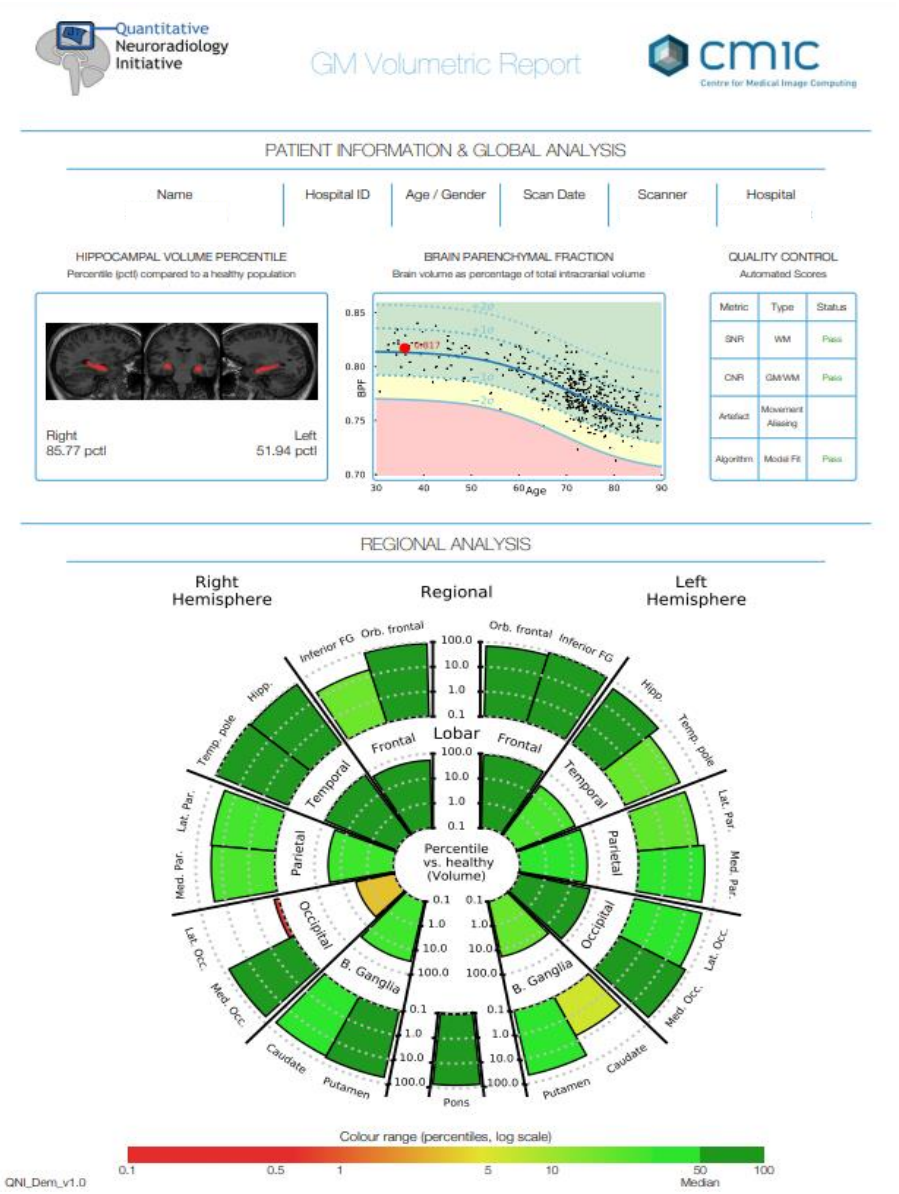


Figure 4-3. Updated quantitative volumetric report.

4.5.2 User experience insights

Changes made to the quantitative report were devised based on the outcomes of the clinical accuracy study, with a view to increasing usability and interpretability of the quantitative information it contains. To directly test how the report is used and interpreted, I conducted a focused user insight assessment. This was done using cases from the clinical accuracy study which were re-processed to generate an updated report. The exercise involved two radiologists, users 1 and 2, one of whom was a neuroradiology consultant and the other who was a specialist trainee completing higher training in neuroradiology. They had not seen the quantitative reports before and received no instructions on how to interpret them. I set up a user workstation with the randomly selected cases and loaded the new quantitative report next to each study. I instructed the users to assess the cases in the way they would usually but to incorporate the quantitative report into their workflow. They were asked to talk through their thinking for each step of the assessment. These 'think-aloud' exercises were followed by some direct questions which targeted specific areas of the report. The full transcripts of these sessions can be found in section 4.7.

4.5.3 User experience outcomes

Performing these user experience interviews provided useful insights into how the quantitative reports are interpreted by target users who have not been exposed to them or received training in their use. A very quick learning effect was observed for both raters, with many elements of the report having been figured out independently within the first or second case.

It was very clear that the raters were far more reliant on the graphical elements and colour scales used in the report than the numerical values and scales provided. For example, the users were not clear on the concept of a logarithmic scale and did not find it easy to assimilate this concept with the corresponding colour scale, however this did not stop them from constructing an accurate interpretation of the colour scale itself. Similarly for the BPF graph, the raters were reliant on the colour bands and the mathematical concepts were secondary, but this did not lead them to draw incorrect conclusions. When they

revisited the axis labels of the BPF graph and log scales, they were able to consolidate the mathematical concepts somewhat. This speaks to the user wanting to ‘dive in’ to using the report in the most accessible way, through use of colours and shapes, highlighting that some structured training to underpin the quantitative concepts that lie behind these presentation techniques would be beneficial before a reporting tool is introduced into a clinical setting.

It was interesting to observe how the users integrated the report findings into their routine assessment. They were not strict in whether they looked at the images or the report first, but they usually made a clear distinction between their visual assessment and the information contained in the report, and were often explicitly articulating whether the quantitative information supported or undermined their visual assessment. There were no cases in which their assessment changed due to the quantitative report, rather it consolidated their visual interpretation. As these were two raters with substantial neuroradiology experience they did not display over-reliance on the report that may be encountered with less experienced raters.

4.6 Plans for a future clinical accuracy study

4.6.1 Reflecting on the completed clinical accuracy study

The clinical validation study that was detailed in section 4.4 of this chapter involved nine raters – three each of neuroradiology consultants, trainees, and non-clinical image analysts – who used a website platform to assess the T1-weighted brain MRI scans of 45 subjects who were evenly split between AD, FTD and subjective memory complaints. The nine raters assessed the scans with and without a quantitative volumetric report available, in a randomly generated order.

The main outcomes of the study were:

- The quantitative report improved sensitivity for detecting neurodegeneration across all raters combined ($p=0.015$).
- Consultant neuroradiologists' assessment accuracy ($p=0.02$) and agreement with the gold standard (kappa scores, $p=0.038$) significantly improved with the use of quantitative reports.
- Inter-rater agreement improved from 'good' to 'excellent'.

However, several issues were encountered and there were some unexpected outcomes, which are important to reflect on and address for a potential further clinical accuracy study. Statistical power was limited by the number of raters who participated, with most results reaching trend-level significance despite moderate to large effect sizes. Disparity was recorded in the performance of raters within the non-clinical image analyst group, and small rater numbers per rater group meant that it was difficult to show a clear benefit for most outcomes. It is possible that there was an element of selection bias affecting the subject cohort. Group numbers between AD, FTD and subjective memory complaints were roughly equal. Raters were not primed to expect an even split, and since this does not reflect true prevalence, may have been a source of confusion for some raters. Additionally, there was increased confidence in incorrect diagnoses when raters had the quantitative report available, which may

suggest an over-reliance on the report and/or misinterpretation of its contents by some raters.

Actions undertaken following the clinical accuracy study included making aesthetic changes to the quantitative report, to include a clearer colour coding system and streamlined use of numbered scales, and undertaking one-to-one simulations and interviews with radiologists while using the updated reports, as has been described earlier in this section.

4.6.2 Planning a future clinical validation study

4.6.2.1 Aims and design of a future clinical validation study

A future clinical validation study would aim to build upon the results found in the pilot study, by performing a larger study involving more raters to reach higher statistical power. It would allow for the updated report format to be tested.

In designing this future study, several design points are important. The setting should reflect the radiologist's normal reporting environment as closely as possible, with the automated report displayed alongside the imaging series. Assessment of radiologists' accuracy, confidence, and reporting efficiency may all be measured. Cases should be presented in a randomised order, twice, with and without the quantitative report being available. The case mix should include a varying degree of structural pathology, from clearly pathological to more subtle changes, as well as normal control subjects. Including a range of severity in the case mix is valuable in discerning whether the added quantitative information is most impactful where the pathology is subtle or unclear. The case mix should also aim to reflect the normal incidence of cases and the commonest patterns of atrophy that are routinely encountered in standard radiology practice in a memory clinic setting. The pathology of each case should be established to the best available gold standard, depending on the condition in question (for example, CSF analysis and neuropsychiatric profile in the case of AD and FTD) (Table 4-7).

Table 4-7. Case selection criteria. The criteria in bold are essential requirements for a case to be selected. bv=behavioural variant; PPA=primary progressive aphasia; DLB=dementia with Lewy bodies; PSP=progressive supranuclear palsy.

Designation	***No. cases***	***Clinical criteria***	***Laboratory marker***	***Radiology marker***
Subjective memory complaints	15	**MMSE ≥ 26** **Clinical assessment has excluded dementia** **Normal cognition confirmed at follow-up**	Normal CSF	MRI: atrophy commensurate with age; no specific pattern suggestive of pathology
Amnestic MCI / Prodromal AD	5	**MMSE 21-26** Change from normal cognition, but not meeting dementia criteria Objective impairment in one or more cognitive domains Preservation of independence in functional abilities **Went on to have confirmed AD at follow-up**	**Abnormal CSF ratio of tau/Aβ**	MRI: a good spectrum - within normal limits/MTA/PCA
AD	10	**MMSE ≤ 24** **Interference with function in usual activities of daily living** Cognitive impairment in minimum 2 domains	**Abnormal CSF ratio of tau/Aβ**	MRI: MTA or PCA
FTD	6	**Clinically confirmed with extended follow-up or genetically** bvFTD: disinhibition, apathy/inertia, loss of sympathy/empathy, compulsive behaviours, hyperorality, dysexecutive neuropsychologic profile PPA: language difficulty most prominent feature and causing impaired daily living, other dx excluded	**Normal CSF ratio of tau/Aβ**	MRI: Frontal and/or temporal lobe atrophy
DLB	5	**Clinically confirmed:** Core clinical features: fluctuating cognition, visual hallucinations, REM sleep disorder, at least 1 parkinsonian feature Supportive clinical features: sensitivity to antipsychotics, postural instability, recurrent falls, syncope, autonomic dysfunction, hypersomnia, hyposmia		MRI: generalised cortical atrophy (more frontal and parietotemporal regions); relative focal atrophy of midbrain, hypothalamus. Relative preservation of medial temporal lobe structures Positive ^{123}I-FP-CIT SPECT (DaTscan)

PSP	4	**Clinically confirmed:** Early postural instability, falls, oculomotor deficits (vertical gaze palsy), akinesia, frontal lobe impairments (speech & language, behaviour), lack of response to levodopa	MRI: **midbrain atrophy**

It would be useful to include several levels of expertise, ranging from experienced sub-specialty radiologists to general radiologists, and radiology trainees. This would provide insight into how the quantitative report interacts with experience level and whether the report affects inter-rater agreement between and within these groups. Rater groups could be defined as consultant neuroradiologists who work in specialist centres and have experience leading dementia MDT meetings; consultant general radiologists who work in general hospital settings but have an interest in brain MRI; and trainees undertaking their specialist training in neuroradiology.

Based on the effect sizes for assessment accuracy from the pilot study, power calculations show that 40 raters would provide a 90% chance of observing a positive effect. Rater recruitment should be from multiple centres. Aiming to recruit raters from a minimum of eight centres would allow for a mix of working practices and prior experience to be represented.

Improved rater instructions are required prior to completing the assessment exercise. This could include a video or animation explaining the format of the quantitative report in more detail. It may also be useful to include some practice cases that raters could perform prior to the real task. An important point to highlight to raters in the instructions is that cases have been selected for dementia-related volumetric abnormality that is able to be assessed on T1-weighted imaging, and not for white matter pathology or any other incidental comorbidities.

A potential extension would be to try to analyse quantitative report engagement metrics, for example by using a survey that is embedded into the rating exercise, thereby gathering real-time feedback, however this would

require very careful planning and design, perhaps guided by a user experience or human computer interaction expert.

4.6.2.2 Power calculations

Based on the results of the clinical validation study carried out, it is possible to calculate that the following sample size estimations to help inform future studies. To achieve an 80%, 90%, and 95% chance of observing a positive effect, 30, 40, and 45 raters would be required to participate in a future study, respectively.

Power assessments using the results of the pilot study have shown that increasing the number of raters would be more effective than substantially increasing the number of cases that they are asked to rate. Increasing the number of scans to be assessed would afford some increased ability to sensitively assess effects of the report within each rater. While this is very important, diminishing returns would be observed with each scan added above around 50. As it has already been shown that there are relatively large effects sizes of the quantitative report, a future study would aim to assess the size and consistency of those effects across a larger pool of raters so that more confident inferences can be made about the effect of the report at the population level.

4.6.2.3 The case cohort

This should aim to strike a balance between representing the typical case mix a radiologist would expect to encounter during a reporting session, thereby providing ecological validity, and ensuring that group numbers are able to provide statistically meaningful experimental sensitivity or construct validity. For example, if ecological validity alone was considered, then the number of FTD cases included should be small. However, the implications of this would be that one mis-rated FTD scan would lead to large changes in accuracy for FTD assessment. If construct validity is to be favoured, raters should be instructed not to take population prevalence base rates into account, at the expense of achieving a true representation of the real-world scenario.

Ecological validity is deduced from population prevalence. Expert consensus suggests that in the UK approximately 62% of dementia is due to Alzheimer's disease, 17% to cerebrovascular disease, 10% to mixed aetiologies, 4% to dementia with Lewy bodies, 2% to Parkinson's disease dementia, 2% to frontotemporal dementia and 3% to other causes (Prince et al. 2014). However, that is different to the case mix of those who present to memory clinics, which includes people who do not end up with a dementia diagnosis.

An NHS England audit of London memory services carried out in 2019 included data from nine memory clinics across London (Cook, Souris, and Isaacs 2019). Of 455 individuals that were seen in these clinics 63% received a diagnosis of dementia, ranging from 49-81% per service. The subtype diagnosis given in those individuals older than 65 years was: AD - 42%, vascular dementia - 18%, mixed dementia - 22%, and other dementias - 13% (Figure 4-4).

Dementia subtypes (aged 65 and over):

	London 2016	Service variation 2016	London 2019	Service variation 2019
Alzheimer's Disease	54%	25% to 77%	42%	30% to 79%
Vascular dementia	10%	3% to 22%	18%	3% to 33%
Mixed dementia	20%	6% to 31%	22%	5% to 36%
Unspecified dementia	12%	0% to 26%	13%	0% to 23%

Table 1 Dementia Subtypes (aged 65 and over) 2016 and 2019

Figure 4-4. Snapshot from Cook et al. 2019, an audit of nine memory clinics in London for the years 2016 and 2019.

An ecologically valid study should more closely mirror the memory clinic prevalence figures than the overall population prevalence figures, since this is more in line with the case mix that a radiologist would encounter.

Example figures are suggested below, maintaining ecological validity. Time investment by raters needs to be factored in so that task engagement is maintained as much as possible. Raters could be expected to spend approximately five minutes on each case, and each case is rated twice, which allows for some time estimates for completing the assessment exercise.

1. 40 cases: 16 normal aging, 16 AD, 8 other – 6-7 hours.

2. 45 cases: 18 normal aging, 18 AD, 9 other – 7.5 hours.
3. 60 cases: 24 normal aging, 24 AD, 12 other – 10 hours.

The quantitative report only represents information about brain atrophy, derived from the T1-weighted imaging sequence, and does not provide any information about other important aspects of dementia pathology, for example white matter hyperintensities or microhaemorrhages. Therefore it is debateable whether cases with vascular pathology should be expressly included in a clinical accuracy study.

In clinical practice, radiologists will encounter individuals with a wide spectrum of vascular burdens, ranging from irrelevant to being the primary cause of the cognitive problem. Since the aim of a clinical validation is to reflect the radiologist's normal reporting environment as closely as possible, and the ultimate intention is to apply the quantification to all patients seen by dementia clinic consultants, which will include patients with vascular disease, then these cases should be represented.

It may be useful to state in the instructions for raters that there are no purely vascular dementia cases included in the case mix, that some cases may have a vascular component, but that the quantitative report is only providing information for the atrophy element of the scan. However, the relationship between white matter lesion loads and grey matter atrophy should be kept in mind when selecting appropriate representative cases (Lambert et al. 2015).

Within each category, it will be important to include a range of severity of cases, guided by radiological and clinical assessment. Cases should be graded by an expert neuroradiologist for their severity, so that a satisfactory mix is established. Typical atrophy patterns should make up the mainstay of cases. For example, for the AD group, most cases should have an MTA pattern, but some could have a posterior cortical atrophy (PCA) pattern, in line with real-world prevalence in the studied age range.

4.7 Supplementary material

In this section the content of the 'think-aloud' exercises and direct interview questions are transcribed in full.

4.7.1 User 1 responses

4.7.1.1 'Think-aloud' cases

Case 1

This is a 56 year old female, 3T scanner. It looks like there's been some segmentation of hippocampal volumes and it's been plotted on – these are presumably standard deviations. And this is plotted in relation to that. Looking at this top row first, it looks like the volumes are two standard deviations below for this red dot. Just trying to correlate what that red dot is because – not quite sure what these mean – it says right 7.44 and left 1.54 – presumably this is percentile? It seems to have passed all the quality control. And I take it this is regional atrophy left vs right hemisphere. Percentile versus healthy volume… so I guess this is showing that there's predominantly frontoparietal volume loss. Inferior frontal gyri bilaterally, left orbitofrontal, and the lateral and medial parietal. So, I can see how that all correlates with the imaging findings, shows that the left side is more affected that the right visually just having a quick look at it, because the left insula looks much more widened and parietal atrophy seems more prominent – so I can now correlate what the report shows to what I can see structurally. But it's easier to interpret that whilst looking at the imaging, or look at the imaging and then interpret that, than just looking at it straight up, having not seen it before.

Case 2

So looking at the report first – a 58 year old female – again with hippocampal segmentation. Again, I wasn't entirely clear what the red dot represents – whether that's an overall volume for both combined or one individually, I'm presuming it's both combined, because there's only one number on here which is two standard deviations below – oh no hold on – sorry I just read this bit up top – brain volume as a percentage of TIV. So I think I misinterpreted those on the previous so – that's the hippocampal volume compared to a healthy population and that's the overall brain volume and I was just previously interpreting that as the hippocampal volume so again that shows that the brain is two standard deviations below what you'd expect for total intracranial volume. Presumably this is for the distribution of age matched controls?

In terms of regional analysis again showing predominantly parietal atrophy in both hemispheres and I'm just trying to figure out what these bits are – okay so they're scales – and again it looks like it's further divided into lobar volumes,

basal ganglia volumes and pontine volume as well. So if I just correlate that again with the scan, according to that what I should expect to see is predominantly parietal and some occipital on the right volume loss. And that looks like it correlates visually with what I'm seeing, right worse than left. I understand it a bit more now, I think I went through it too quickly last time. I didn't take time to understand what each bit represented.

Case 3

This is a 65 year old female – looking at the top row first and getting a feel for the hippocampal volumes percentiles – which are similar bilaterally – and the overall brain parenchymal fraction – again appears reduced compared to what presumably represents a kind of population, potentially age-matched – oh so the age is down here – okay fine that's also now a bit more clear that the age is at the bottom. The patterns of atrophy show predominantly orbitofrontal, temporal pole and lateral parietal of the left and right. So I guess the next step is looking at this and putting it all together into whether it fits a particular pattern of atrophy that might kind of give a particular diagnosis – because normally I guess you'd look at the imaging and whether there's particular regional atrophy and whether that would point you towards a specific diagnosis. But I think what I've found most useful is looking at that in conjunction with the imaging and then correlating the two together. The only other thing is that this is all based on T1 – I'd obviously also look for other sequences as well looking at things like FLAIR for white matter lesions and vascular disease burden, T2 for strategic infarcts, and SWI for any microbleeds, so then you could combine the segmented volumetric analysis in with that, and that would give you more of a broader picture as to what was going on.

4.7.1.2 Direct questions

1. How do you interpret the hippocampal percentiles that are given?

That's a good question. So it's bilaterally around 13.6 percentile. But I'm not sure how that compares in terms of the normal variation that you might expect. I mean this is compared to a healthy population, but is that age-matched or not? That's what I'd like to know. It might be more clear if it was in a format similar to the BPF which compared by age and how it deviates by standard deviation. So I'm not entirely sure how to interpret that in isolation.

2. How do you interpret the colours?

It's useful when I looked at it in the context of the imaging. Just looking at it as it is, I'm just looking at it as a very basic red/yellow/green colour scheme, red

showing the areas where the atrophy is most marked, and using that to generate a picture of what I might expect to see on the imaging. I'm not using it particularly as – I guess it looks like a log scale? Or percentiles? – but I'm not looking at the numbers specifically, just looking at the colours to see if there's a regional distribution of atrophy, and then comparing it with what I see on the imaging. More as a visual.

3. Have you used the colour scale at the bottom?

No, I actually hadn't looked at that. Colour range – percentile – log scale. So it looks like there is a lot more colour range variation for the lower 50% compared to the top 50%. But I'm not entirely sure how to interpret that either. I'd have to look at it a bit more, read around it a bit more, how the percentiles apply and how to interpret those. I'm quite neutral about it. I think it might be useful but I would need to know more about how to apply that in the context of this graph.

4. How do you interpret the bullseye plot inner and outer circles?

I think this [inner circle] overall represents the lobes and the outer circle is looking at specific regions of that lobe. So it gives you the frontal lobe overall and then a specific gyrus and how that area is contributing to the overall volume loss. Only covers specific gyri, maybe that's because they're more implicated in the dementia process, that's why they've been picked out. But again I'd probably look at that in the context of the imaging rather than in isolation. You can always go back to the imaging and see if other areas are affected that haven't been shown on the plot.

4.7.2 User 2 responses

4.7.2.1 'Think aloud' cases

Case 1

In the top left you've got a hippocampal volume relative to a healthy population so on the right is a 7.4 percentile and on the left a 1.5 percentile. And I think that's obviously at the low end of normal percentiles, so at the bottom end. And then the brain parenchymal fraction I suppose is a surrogate for the global cortical atrophy scale so we're looking at where in terms of the global atrophy or the brain volume of this patient as per an expected intracranial volume and again they're in the bottom 25, bottom fraction, so they've got an abnormal brain parenchymal fraction. The quality control I suppose rates the quality of the scan in terms of its signal to noise, contrast to noise, the amount of artefact, and it passes on all of those fronts. And then the regional analysis, okay so

this gives a regional assessment of where the degree of volume shrinkage compared to a healthy person. Okay so for a 56 year old female there's quite marked cortical atrophy which is most marked in the parietal and temporal regions but also affecting the orbital-frontal regions on the left hand side. The hippocampal volumes are reduced on both sides more severe on the left-hand side as shown in the hippocampal volume percentile. So in a patient of this age, you would look for more posterior if you were thinking of Alzheimer's or presenile Alzheimer's you'd look for a more posterior atrophy pattern so let's have a look at that. So yes that tends to fit, there's posterior parietal atrophy involving the precuneal areas of the brain, more so than the frontal regions, so yes this would probably do for presenile onset of Alzheimer's.

Case 2

60 year old female patient. Doesn't tell us any clinical details. [Scrolls though the scan]. Okay so there's abnormal brain volume loss particularly affecting the left cerebral hemisphere particularly affecting the frontal lobe and temporal lobe for example the parasylvian regions are abnormally widened. Looking at the regional analysis that seems to fit so inferior frontal gyrus and also shrinkage in the temporal lobe. At first glance the hippocampal volumes don't appear to be too bad. I'm just trying to find the colour map. Basal ganglia occipital parietal temporal okay. But the left hippocampal volume on the regional analysis is reduced. Let's have a look. Okay there is widening of the sylvian fissure as we noted on the axial. On the coronal it's on the left side more than on the right side. The architecture of the hippocampus is quite difficult to assess on the left side but it does look asymmetrical and you can also see that the choroid fissure is open as well. And it better demonstrated the degree of sylvian fissure widening as well and posterior frontal atrophy. So I suppose in a patient of this age depending on what the clinical picture was I'd be thinking of an FTD pattern of dementia. And just seeing if that correlates…yes that's about right.

Case 3

Okay so this is a 65 year old female scanned on 3T, just having a quick look at the hippocampal volume percentiles, so bilaterally they're reduced but appears to be fairly symmetrical in the reduction on both sides. The brain parenchymal fraction is reduced to below the sort of – I don't know what that centile is supposed to mean – but the overall fraction is reduced for age. The study meets the quality control standards. Looking at the regional analysis we'd expect to see atrophy in both temporal poles, on the left side predominantly as well as bilateral lateral parietal volume loss. The rest of the brain in the regional analysis appears okay. So I just want to go through the axials and give myself a general overlook [scrolls though the scan]. So left temporal pole is very… so there is asymmetry at the left temporal pole with increased widening and parenchymal volume loss. There's widening of the

lateral sulcus of the sylvian fissure on the left. Frontal volumes look reasonably well preserved. And quick look at the hippocampi they don't look too affected so I'll have a look at those on the coronal. Okay they are affected. So the patient is 65 so just in that indeterminate region if you are using the MTA scoring system. But yes there's opening of the choroid and sylvian fissure on both sides and there's at least an MTA 2 on both sides, possible MTA 3 looking at the hippocampi on the regional analysis, supposedly they're okay. I would say these are MTA 2 on both sides. But yes they're symmetrical. I think with the anterior temporal pole atrophy on the left you'd probably be thinking again an FTD and probably a semantic type FTD depending on the clinical findings and the neurological assessment.

Case 4

So a 77 year old female scanned on the 3T. Quality control is fine. At a glance the parenchymal fraction adjusted for age looks fine and the hippocampal volumes look reduced on the left but within the 50th centile so an average centile on the right. So looking at the regional analysis there looks to be some slight reduction in the medial parietal volume on the right but otherwise nothing particularly striking. So the colour range would make it… I don't really understand this log… oh it's a log scale okay fine so… it's not particularly intuitive the log scale but towards the lower end of the centile range between 2 and 3 on the log scale, so let's have a look at the scan… [scrolls though the scan]. It's not a great quality scan. Quite a lot of motion. Some lacunar infarcts in the very deep perforator territories. There isn't any obvious asymmetry even in the medial temporal regions but I'll just assess that on the coronal. So I think for the age the mesial temporal structures and the hippocampi in particular are passable, I wouldn't say that they were pathologically atrophied. Just looking… It's probably symmetrical posterior parietal atrophy but again I would probably pass this for age and according to the brain parenchymal fraction the brain volume is relatively well preserved. So yeah I don't think there is an obvious neurodegenerative disorder here based on this but obviously you'd have to marry that up with the clinical findings and the mesial temporal structures look okay for age.

4.7.2.2 Direct questions

1. Can you tell me how you interpret the percentiles for the hippocampal volumes?

So it says compared to a healthy population… so they've obviously taken an age adjusted standard and taken the volumes and segmented their hippocampi and compared them to that. So I'm looking at it and I'm seeing if firstly there's any reduction. I don't really know about this scale, as in what you'd call a normal centile but obviously the lower the centile the more chance

that that's pathological for age, and then the second thing I'm looking at is if there's a big discrepancy between the two sides. So here there's a 25 percentile difference between the left and the right so that's how I'm interpreting that.

2. Can you tell me about the BPF graph?

Okay so this I suppose, just looking at the brain volume as a percentage of the total intracranial volume and again adjusted for age, so it's like a normogram, and you're seeing where again segmental analysis of the entire brain volume, where it roughly falls on the normogram. So you can see in this patient it's within the green zone which is taken as between 0.78 and 0.85 so essentially that fraction as a percentage of the total intracranial volume so 78-85% is what you'd expect in a normal fraction for the population for age. The blue dotted lines are a bit distracting. I'm guessing the solid blue line equates to the line of best fit through that scatter plot. And then you've got probably standard deviations away from that mean? That's probably how I would interpret it. So positive and negative away from the mean. So you can see the cluster of people closest to the line represents the mean and that's the standard deviation from the mean. But I think depending on how long you have to look at the study, it would be useful to have a look at this to start with and to try to make sense of it because I think if you're trying to do this quickly that's not particularly intuitive to analyse.

[Unprompted] The QC is quite straightforward, it's completely intuitive, and it's useful because obviously you don't want to extrapolate more from the scan than the quality of the scan, so that's a good feature to have. And especially if you're comparing between centres with different imaging parameters, different scanners, then it's useful to have that as well.

3. How do you interpret the regional analysis?

That's actually really useful. Again it takes a little bit of time to work out the traffic lights and the way it's laid out. Again when you look at it the first time you probably think it's a bit over the top, a bit too much data but then the more cases I did the more I looked at that to start with to get an idea of the overall picture of what I was expecting to then see on the scan. I probably used the colour coding more than I actually used the numerical data. So the logarithmic scale I'm not particularly statistically orientated but someone would probably just have to explain that before I used that just to make sure I understand it. I don't really understand the logarithmic scale. Is there a reason why it couldn't be just a linear numerical scale? So I think the colour coding element is really good, really intuitive and helpful for correlation between the structural findings

on the scan and the data. Inner circle is the lobes of the brain and the outer circle is segmented into different parts of that lobe. So for example you can see the temporal lobe overall and then also taking into account those two other segments of the temporal lobe, and then if you go further out of the tree, because it's a tree-and-branch sort of thing, you see what accounts for the reduction in the temporal lobe volume. So if you take the basal ganglia, it's a green, so it's between 10 and 50 on the scale. But you can see that the putamen is essentially completely normal, 100% on the scale but the caudate is between 1- 10 so probably about 5, so taken together, overall they average out to being within the healthy range I suppose.

4. Is there anything that you would change?

Least intuitive is the log scale and all the numbers, because I think it needs to be as user friendly as possible. Obviously you don't want to say that it's not going to be totally user friendly but if you're having to look at all these scales as well as looking at all the rest of it, it's quite busy. And also if you're showing it in an MDT or something, you want to be able to reference to the clinicians where the abnormality is, and I think the colour coding stands out and makes sense so I'd say that and the BPF when you get your eye into it is good as well. But the least intuitive would be the actual log scales being put on the illustration. It doesn't add much to have the scale at the bottom or the numbers going up the middle. Overall I think the traffic lights and regional analysis is quite useful and especially the way it's structured in terms of the way its fanning out, that's quite nice. It gives you a visual representation and it's radiologically orientated so the right is on the left and the left is on the right so the same way that you're interpreting the scan you're interpreting the picture.

5. Quantitative analysis for MRI in multiple sclerosis

5.1 Introduction

5.1.1 Multiple sclerosis (MS)

MS is an autoimmune disorder affecting of the central nervous system (CNS). It commonly first presents in young adults aged 20 to 30, and is characterised by inflammation, causing demyelination, and axonal loss, which leads to neurodegeneration (McGinley, Goldschmidt, and Rae-Grant 2021; Sand 2015). Its aetiology is not fully understood but it is thought to result from complex interactions between genetic and environmental or lifestyle factors. There is higher prevalence of MS within families, supporting a genetic element, the commonest genetic locus being a major histocompatibility complex, HLA-DRB1. Prevalence is also higher at higher latitudes worldwide, reaching 300 per 100,000 people, thought to be linked to lower sun exposure and therefore lower vitamin D levels in these geographical regions.

MS is characterised by lesions throughout the CNS, typically affecting myelinated regions of white matter (WM). Lesions arise through inflammation and damage to the myelin that surrounds axons. While WM lesions are the typical hallmark of MS, lesions also affect the grey matter (GM). Axonal damage leading to neuronal loss has in recent years been recognised to occur from the early stages of disease (Nygaard et al. 2015; Pérez-Miralles et al. 2013).

Diagnosis of MS relies on a combination of clinical, imaging and laboratory features that must fulfil specific criteria in a patient where there is a high a priori suspicion of a positive diagnosis and other inflammatory neurological conditions that can mimic MS have been excluded (Thompson et al. 2018). A patient's initial symptomatic demyelinating event is referred to as clinically isolated syndrome (CIS), where the patient experiences symptoms typical of MS, commonly involving the optic nerve, brainstem or spinal cord, and MRI examination demonstrates lesions consistent with MS, but the pattern does not fully meet the MS classification criteria. These patients have a high probability of converting to relapsing-remitting MS in the future (Kappos et al. 2007; Kuhle et al. 2015). Conversely, a radiologically isolated syndrome (RIS) is an uncommon incidental presentation defined by the presence of classical

MS features on MRI examination in the absence of any clinical syndrome, where no possible alternative preferred diagnosis is evident (De Stefano et al. 2018). A large retrospective study showed that 34% of individuals with RIS experience their first clinical event consistent with CIS or MS within five years, and analysis suggested that younger age, male gender and spinal cord lesions are the most significant predictors for an individual with a RIS experiencing a future clinical event (Okuda et al. 2014).

Most patients with MS have the relapsing-remitting form of disease, referred to as RRMS. It is defined by acute clinical exacerbations interspaced with periods of recovery, which is either complete or incomplete, from which disability tends to accrue incrementally over time. Diagnosis is made based on a combination of the clinical presentation of typical symptoms and signs, and consistent imaging features.

As well as ruling out any differential diagnoses as being more likely, the diagnosis of MS also must meet the criteria of dissemination in space (DIS) and dissemination in time (DIT). DIS is a requirement that lesions are detected in at least two distinct areas of the CNS that are typically affected by MS. This may be confirmed clinically by two demyelinating events affecting different neurological sites, or radiologically, by confirming the presence of typical lesions on MRI in at least two anatomical territories: periventricular, juxtacortical, infratentorial, and spinal cord. The DIT criteria requires that CNS lesions can be shown to have developed over time, to reduce potential misdiagnosis of monophasic inflammatory CNS events. Similarly to DIS, this criteria can be met clinically or radiologically. Only lesions that are active enhance with gadolinium, so a combination of enhancing and non-enhancing lesions on a single MRI scan indicates DIT. Two serial scans may show the appearance of new lesions with reference to the initial scan and also fulfil DIT criteria.

While the relapsing-remitting form of MS is the most common, there are also important progressive phenotypes. Primary progressive MS, PPMS, is characterised by the unremitting progression of symptoms and decline in neurological function from the time of diagnosis, and affects about 15% of

patients with MS (Miller and Leary 2007). Secondary progressive MS, SPMS, describes established progression of decline in neurological function following an initially relapsing disease course. After 20 years of RRMS around 40% of patients will transition to SPMS (Rovaris et al. 2006).

In addition to the phenotypic descriptions of MS disease course, which are limited and do not sufficiently reflect the underlying pathophysiology, terminology expressing the presence or absence of disease activity has evolved. Disease activity encompasses relapses, progression between phenotypes, and MRI evidence of new inflammatory lesions and pathological brain atrophy, representing active neurodegeneration (Lublin et al. 2014). The concept of 'no evidence of disease activity' (NEDA), defined as the absence of relapses, disability progression, and active MRI lesions, is frequently used as an endpoint in clinical trials for disease-modifying MS treatments.

Basing MS classification on its classical clinical phenotypes is increasingly acknowledged as insufficient, since they are removed from the underlying pathophysiological phenotypes, demonstrate significant overlap in terms of imaging, clinical and laboratory findings, and transitions between them are difficult to pinpoint accurately. Recently, machine learning techniques have been used to identify common features between patients with MS based on their brain MRI, and the novel MRI-based 'cortex-led', 'normal-appearing WM-led', and 'lesion-led' subtypes have been shown to have distinct patterns of disability progression and responses to treatment (Eshaghi et al. 2021). Developments like this could help to stratify patients more effectively for clinical trials and eventually facilitate treatment personalisation (Gafson, Craner, and Matthews 2017).

5.1.2 Neuroimaging biomarkers of MS using standard of care images

MS pathology is driven by both innate and adaptive immune system, with activation of microglia and macrophages, as well as B and T lymphocytes, which are recruited by CNS-specific target antigens. These mediate demyelination and axonal loss, the latter being the main driver of neurological decline and permanent clinical disability. Conventional MRI techniques can be

used to detect the neuroimaging correlates of inflammation, demyelination and axonal loss.

Typical MS lesions in the brain are ovoid in shape, aligned perpendicular to the lateral ventricles and perivenular (Sand 2015). They appear as areas of high signal intensity on T2-weighted and T2-FLAIR sequences and are iso- to hypo-intense on non-contrast-enhanced T1-weighted MRI. Recent consensus recommendations between European and North American expert advisory groups promote the use of standardised imaging protocols. They recommend that 3T imaging is preferable to 1.5T if available, and core sequence acquisitions should be a three-dimensional (3D) T2-FLAIR sequence, an axial T2-weighted sequence, and a gadolinium contrast-enhanced T1-weighted sequence at baseline (Wattjes et al. 2021). For follow-up imaging, contrast-enhanced T1-weighted imaging is not recommended. Quantification of inflammatory activity in terms of number and volume of MS-typical lesions is central to patient monitoring and clinical trial outcome reporting. While lesions that affect the cortical grey matter have been shown to be an important correlate of long-term disease progression (Haider et al. 2021), they are difficult to detect on standard of care imaging protocols (Geurts et al. 2011).

The neuroimaging correlate of axonal loss is the observation of established reduced brain volume that is more accelerated than due to an individual's age alone. Brain volume loss in MS has been shown to occur at more than 0.4% per year (De Stefano et al. 2016), compared to an expected 0.05-0.3% volume loss observed in healthy controls depending on their age (Battaglini et al. 2019). Even faster rates of atrophy may be observed in the deep grey matter (Bishop et al. 2017; Eshaghi et al. 2018). Brain volume loss on MRI has been shown to correlate with actual histological tissue loss (Popescu et al. 2016) as well as with short- and long-term measures of clinical disability (Losseff et al. 1996; Scalfari et al. 2018). Atrophy of the grey matter seems to correlate with clinical disability and cognitive impairment more than whole brain or WM atrophy (De Stefano et al. 2003), and has been shown to occur from disease onset in association with acute inflammation (Chard et al. 2002). While it is not currently utilised in disease classification or diagnostic criteria, the importance of brain atrophy as a prognostic factor is recognised in the research field

(Sastre-Garriga, Pareto, and Rovira 2017). Imaging standardisation and improved quantification methodology as well as technical and clinical validation at the individual patient level could facilitate its implementation as a useful clinical tool (Rocca et al. 2017).

5.1.3 Quantification of MS neuroimaging biomarkers

Volumetric quantification techniques measuring MS lesion load and brain volume have been developed in the research setting and have been used in clinical trials, where image acquisition is somewhat standardised and quality controlled, and where multiple image contrasts are available (Danelakis, Theoharis, and Verganelakis 2018; Lindig et al. 2018). Lesion segmentation techniques generally rely on the availability of 3D, multi-contrast source image data, i.e. requiring both 3D T1- and typically T2-FLAIR-weighted images (de Boer et al. 2009; Simões et al. 2013). Classical brain tissue segmentation typically requires a 3D T1-weighted image dataset. Tissue segmentation can be affected by the presence of WM lesions, and this can be mitigated against by detecting and correcting for them (Valverde et al. 2015).

Radiologists could benefit from incorporation of automated volumetric lesion and brain volume assessments in their routine clinical workflow. Recent promising developments have been made towards clinically useful solutions that are able to tolerate image quality and acquisition heterogeneity, by using T2-FLAIR to measure central (i.e. periventricular) lesion volume and central atrophy as representative measures of the whole brain status. This has been shown to be effective and reproducible using heterogenous clinical data including two-dimensional (2D) T2-FLAIR data (Dwyer et al. 2019; Zivadinov et al. 2018). However, clinical centres are increasingly adopting guidelines for a 3D T2-FLAIR protocol, and proprietary tools focus on providing 3D volumetric quantification.

Translating group-level findings to the individual patient is challenging. The provision of reference data against which an individual's quantitative results can be compared is important, as it provides context to the output values. In one study, comparison of individual MS patients' brain volumes to a large cross-sectional reference dataset showed that patients with a low brain volume

compared to the reference were more than two times more likely than other patients to develop disability progression over the next two years (Sormani et al. 2017). While studies have shown that global brain volume measures are associated with clinical outcomes in MS patient populations, it is more challenging to show associations with specific regional cortical and deep GM atrophy patterns that are known to occur in MS (Pichler et al. 2016). This may be due to limitations in segmentation techniques and underlying variability of clinical data (Sastre-Garriga et al. 2020).

5.1.4 Commercially available quantitative imaging reports for MS

Several proprietary software tools have been developed specifically for quantification of MRI biomarkers for clinical use in MS patients. These tools have received certification for clinical use from entities like CE and FDA, however they are generally sparse in their technical and clinical validation in comparison to established academic research tools.

The icobrain MS tool is a quantitative report provided by the Belgian company Icometrix (icometrix 2021). It combines assessment of WM lesions and brain volume, using dual T1 and T2-FLAIR input (Jain et al. 2015). The report provides snapshots of WM lesion and cortical GM segmentations, lesion volume quantification and distribution in the periventricular, juxtacortical, deep white matter and infratentorial regions, and whole brain and GM volumes, also expressed as percentiles in the context of a normative reference population, details of which are not publicly available. It includes a longitudinal element providing annualised percentage volume change and whether there are new or enlarging lesions.

The proprietary method was included in an independent technical validation study against established research methods applied to an MS population, where substantial differences were shown to exist between methods and technical issues including sensitivity to scanner upgrade were highlighted as important issues that may undermine the tool's direct clinical applicability (Steenwijk et al. 2017). Beyond technical comparisons, independent clinical validation studies involving neuroradiologists have not yet been published.

LesionQuant is another example of a commercial solution, offered by Cortechs.ai (cortechs.ai 2021) as an extension to their NeuroQuant brain volume quantification tool (Brewer et al. 2009). It also presents lesion quantification metrics by region and highlights whether lesions are new, expanding or shrinking in a longitudinal comparison. It does not provide reference data for lesion quantification measurements; this is only provided for the brain regional volumetric section of the report. Like icobrain MS, it also requires combined T1-weighted and FLAIR input, and its proprietary lesion segmentation method is not publicly accessible. It has been assessed in a clinical validation where its results were compared to a neuroradiologist's visual scoring in a group of 56 patients with MS with good correlation but with a tendency to underestimate lesion counts (Brune et al. 2020). Standalone head-to-head comparisons like this do not replicate the clinical scenario and underscore the importance of assessing quantitative report information combined with and as an adjunct to the radiologist's visual assessment.

5.1.5 Developing a FLAIR-only quantification pipeline

As mentioned, while several proprietary solutions exist that perform lesion and brain volume quantification, these require 3D T1-weighted images as well as T2-FLAIR images. In the standard clinical routine, a 3D T1-weighted scan is not routinely performed and not recommended by current guidelines, apart from a contrast-enhanced diagnostic study (Schmierer et al. 2019). While a possible solution to this would be to advocate for changing recommended acquisition protocols to include a 3D T1-weighted sequence, this adds to clinical workloads, and it is therefore worthwhile exploring the value of T2-FLAIR-only quantification. Proprietary solutions are also variable in the information they offer, some providing only lesion-related data, and variable reference data which often does not include MS-population reference data. It is also difficult to determine how these solutions have been validated for clinical application, and the gold standard they have been assessed against (Wilkinson and van Boxtel 2019). These issues present a considerable translational barrier for integration of well-validated and valuable quantitative techniques for clinical MS imaging. The differences between lesion and brain

volume quantification in the research and clinical settings is illustrated in Figure 5-1.

The aim of the work in this chapter is to address an aspect of this important translational challenge, that of performing automated lesion and tissue segmentation using a single T2-FLAIR input. The performance of T2-FLAIR-only quantification will be compared to a conventional multi-sequence approach. Additionally, the performance at different field strengths will be investigated by using a multicentre dataset of subjects with CIS who have been scanned at 1.5T and 3T within the same week.

This work will aim to establish whether T2-FLAIR-only lesion and brain segmentation introduces more variability compared to conventional techniques, with the hypothesis set that T2-FLAIR-only segmentation will achieve comparable results.

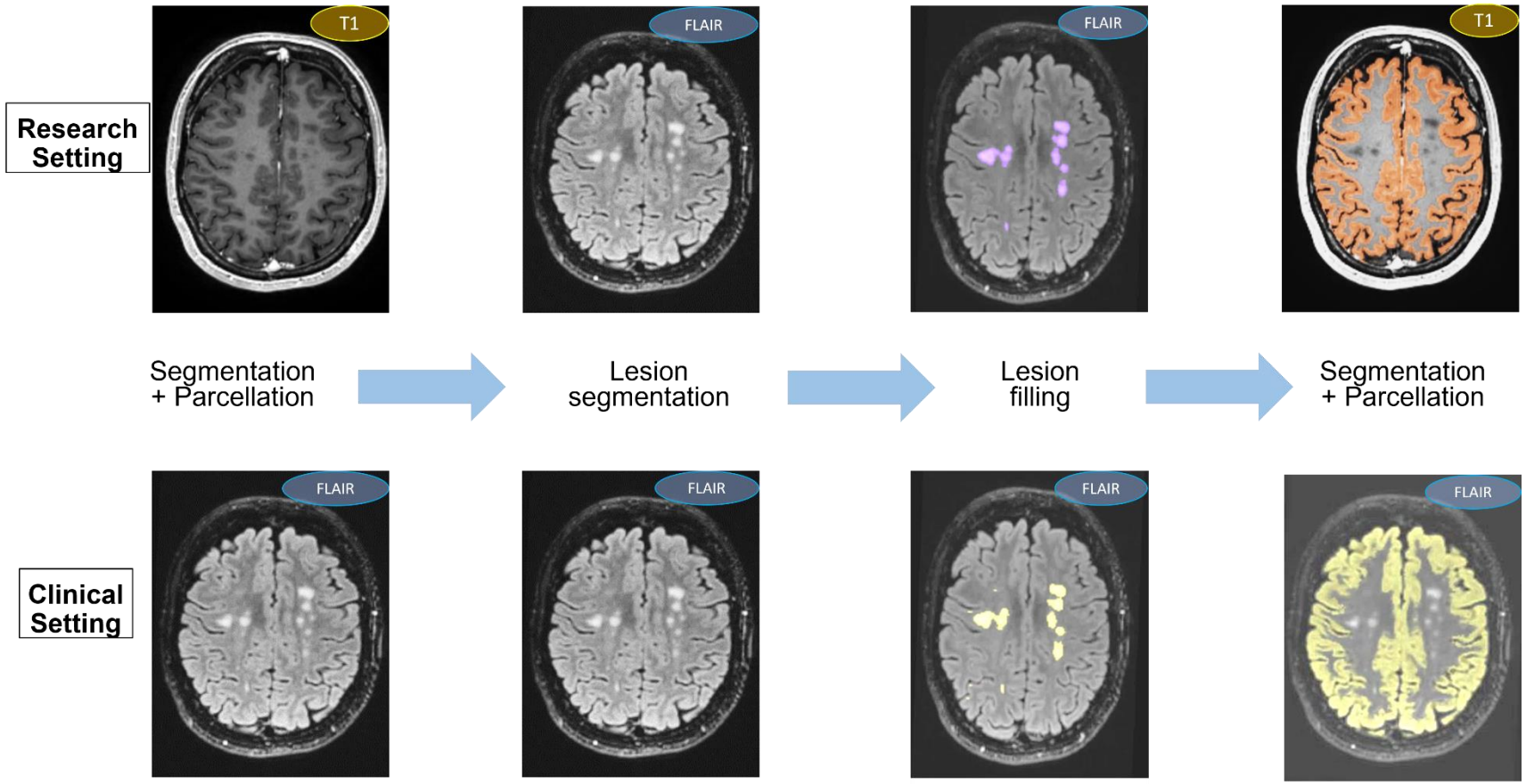


Figure 5-1. Illustration of the processing steps required for lesion and brain tissue segmentation in the research setting, where a T1 sequence is available, versus the clinical setting, where all steps must be performed on a FLAIR sequence.

5.2 Methods

5.2.1 Dataset

The dataset used, which has previously been described elsewhere (Hagens et al. 2018), consists of subjects recruited between 2013 and 2015 from six different clinical MS centres across Europe, who are members of the Magnetic Resonance Imaging in Multiple Sclerosis (MAGNIMS) network (www.magnims.eu, MAGNIMS 2021). For this study, a subset of 66 subjects with CIS were used.

The inclusion criteria for CIS subjects were defined according to the international panel on MS diagnosis (Polman et al. 2011), and subjects had no other immunological, vascular or oncological previous medical history. The study was approved by local review boards and all participants gave their written informed consent to participate.

5.2.2 MRI acquisition

MRI scanning was performed on 1.5T and 3T scanners within the same week. Scanning parameters were applied in line with established MAGNIMS guidelines (Wattjes et al. 2015), using a multi-sequence scanner optimised acquisition protocol, which included in particular an isotropic gradient echo 3D T1-weighted and 3D turbo spin echo T2-FLAIR sequence. Table 5-1 details the acquisition parameters for each centre and field strength.

Table 5-1. MRI sequence parameters by centre, for 1.5T and 3T.

MAGNIMS Centre	*1.5T*				*3T*			
	Vendor	Parameter	3D T1	3D FLAIR	Vendor	Parameter	3D T1	3D FLAIR
VU University Medical Center Amsterdam	GE Signa HDxt	Type	GRE	TSE	GE Discovery MR750	Type	GRE	TSE
		Slice orientation	Sag	Sag		Slice orientation	Sag	Sag
		Measured voxel size (mm)	1.0x1.0x1.0	1.4x1.4x1.2		Measured voxel size (mm)	1.0x1.0x1.0	1.1x1.1x1.2
		TR (ms)	12.4	6500		TR (ms)	8.2	8000
		TE (ms)	5.2	115		TE (ms)	3.2	130
		Flip angle (degrees)	12			Flip angle (degrees)	12	
		Turbo factor		191		Turbo factor		230
		Inversion times (ms)	450	1994		Inversion times (ms)	450	2340
University Hospital Basel	Siemens Avanto	Type	GRE	TSE	Siemens Verio	Type	GRE	TSE
		Slice orientation	Sag	Sag		Slice orientation	Sag	Sag

		Measured voxel size (mm)	1.0x1.0x1.0	1.0x1.0x1.0		Measured voxel size (mm)	1.0x1.0x1.0	1.0x1.0x1.0
		TR (ms)	2700	6000		TR (ms)	1570	5000
		TE (ms)	3.37	352		TE (ms)	2.67	402
		Flip angle (degrees)	8			Flip angle (degrees)	9	
		Turbo factor		141		Turbo factor		141
		Inversion times (ms)	950	2200		Inversion times (ms)	900	1800
St. Josef Hospital Bochum	Siemens Avanto	Type	GRE	TSE	Philips Achieva	Type	GRE	TSE
		Slice orientation	Sag	Sag		Slice orientation	Sag	Sag
		Measured voxel size (mm)	1.1x1.1x1.0	1.0x1.0x1.0		Measured voxel size (mm)	1.1x1.1x1.0	1.0x1.3x1.0
		TR (ms)	10	4800		TR (ms)	10	5000
		TE (ms)	4.6	291		TE (ms)	4.2	354
		Flip angle (degrees)	8			Flip angle (degrees)	15	
		Turbo factor		204		Turbo factor		182
		Inversion times (ms)	1000	1650		Inversion times (ms)	1100	1900
UCL Institute of Neurology London	Siemens Avanto	Type	GRE	TSE	Philips Achieva	Type	GRE	TSE
		Slice orientation	Sag	Sag		Slice orientation	Sag	Sag
		Measured voxel size (mm)	1.0x1.0x1.0	1.0x1.0x1.0		Measured voxel size (mm)	1.0x1.0x1.0	1.2x1.2x1.2
		TR (ms)	1900	6500		TR (ms)	6.9	8000
		TE (ms)	3.37	202		TE (ms)	3.1	388
		Flip angle (degrees)	15			Flip angle (degrees)	8	
		Turbo factor		125		Turbo factor		120
		Inversion times (ms)	1100	2000		Inversion times (ms)	821	2400
Hospital Clínico San Carlos Madrid	GE Signa HDxt	Type	GRE	TSE		Type	GRE	TSE
		Slice orientation	Sag	Sag		Slice orientation	Sag	Sag
		Measured voxel size (mm)	0.98x0.98x1.0	0.98x0.98x1.0		Measured voxel size (mm)	0.98x0.98x1.0	0.98x0.98x1.0
		TR (ms)	10	6000		TR (ms)	10	6000
		TE (ms)	4.2	136		TE (ms)	4.2	135
		Flip angle (degrees)	12			Flip angle (degrees)	12	
		Turbo factor		220		Turbo factor		220
		Inversion times (ms)	450	1837		Inversion times (ms)	450	1840
Sapienza University of Rome	Siemens Avanto	Type	GRE	TSE	Siemens Verio	Type	GRE	TSE
		Slice orientation	Sag	Sag		Slice orientation	Sag	Sag
		Measured voxel size (mm)	1.0x1.0x1.0	1.2x1.2x1.3		Measured voxel size (mm)	1.0x1.0x1.0	1.0x1.0x1.0
		TR (ms)	1900	6500		TR (ms)	1900	5000
		TE (ms)	3.37	202		TE (ms)	2.93	395
		Flip angle (degrees)	15			Flip angle (degrees)	9	
		Turbo factor		125		Turbo factor		141
		Inversion times (ms)	1100	2000		Inversion times (ms)	900	1800

GE=General Electric; Sag=sagittal; GRE=gradient echo; TSE= turbo spin-echo; TE=echo time; TR=repetition time

5.2.3 White matter lesion detection

All scans were read during the original study by Hagens and colleagues by consensus joint reading using a digital workstation (Sectra [Linköping, Sweden] IDS7 version 16.2.28), by three experienced readers in random order, with a minimum interval of two weeks between reading 1.5T and 3T scans. White matter lesions were defined as areas of abnormal WM hyperintensity, consistent with CIS lesions, that were apparent on T2-FLAIR images and were larger than 3mm in diameter. The image readers had

information regarding localisation of clinical signs and symptoms as detected by the neurologist, but were agnostic to subject age, gender and clinical centre.

5.2.4 Manual WM lesion segmentation

To investigate whether automated lesion segmentation resembles segmentation of any white matter hyperintensity, or typical MS lesions more closely, two types of manual segmentation were performed. Rater 1 performed manual segmentation of CIS lesions as guided by the expert consensus labelling described above (referred to in results as ‘manual method 1’). Rater 2 performed manual segmentation on a subset of subjects not guided by the expert consensus labels, and instead included any hyperintensity in the white matter (referred to in results as ‘manual method 2’).

5.2.5 Automated WM lesion segmentation

T1 and T2-FLAIR images were segmented using the Bayesian Method of Model Selection (BaMoS) (Sudre et al. 2015), which is an unsupervised hierarchical model selection framework, which facilitates distinction of different types of expected and unexpected signal intensities following brain parcellation. Standard two sequence input segmentation with T1 and T2-FLAIR images was performed, and then repeated on the same dataset using T2-FLAIR as the only input sequence. This involves fitting a Gaussian mixture model to the data, and optimising the number of components required for each tissue class, in a similar way to the original two sequence method. The output parcellations obtained were post-processed using a database composed solely of T2-FLAIR images for the removal of false positives.

5.2.6 Brain tissue segmentation

Brain segmentation was performed using Geodesic Information Flows (GIF) (Cardoso et al. 2015), a fully automated multi-atlas-based approach which has previously described in detail in chapter 4. This was done in two ways:

1. Using a 3D T1 image database, namely the original GIF database composed of T1 images that were expertly manually labelled, and

2. Using a newly-constructed GIF database which contained both 3D T1 and 3D T2-FLAIR images.

5.2.7 The new multi-modal GIF database

The new database was constructed using co-registered 3D T1 and 3D T2-FLAIR imaged from 100 healthy control subjects derived from the Southall and Brent Revisited 'SABRE' study cohort (Tillin et al. 2012) (age range 46-90 years, mean age 72, 51.1% males).

The acquisition parameters were:

1. 3D sagittal T1 multishot, inversion-prepared gradient echo: repetition time 6.9 ms; echo time 3.1 ms; voxel size 1.0x1.0x1.0 mm^3; and
2. 3D sagittal T2-FLAIR: repetition time 4800 ms; inversion time 1650 ms; echo time 125 ms; voxel size 1.0x1.0x1.0 mm^3.

The original T1 labels were used to automatically segment the new T1 images, and these labels were then propagated to the T2-FLAIR images. Performance of GIF with the original and new image databases was compared by segmenting the CIS cohort's 3D T1 images for direct comparison of the effect of database change. The new GIF database performance was then tested with 3D T1 only, and T2-FLAIR only images from the CIS subject cohort. To assess tissue segmentation performance in the context of high WM lesion loads, a subset analysis of the 10% of cases with the largest lesion volumes was performed.

Prior to segmentation T2-FLAIR images were registered to T1 space to allow for voxel-wise comparisons. Performance was tested with varying degrees of WM lesion inpainting (Chard et al. 2010). This was done using a patch-based method (Prados et al. 2016) to make three comparisons: uncorrected, manual WM lesion filled, and BaMoS outlier filled.

5.2.8 Statistical Analysis

5.2.8.1 WM Lesions

WM lesions were assessed by segmentation method and field strength for 1) absolute lesion volume (median and interquartile range, IQR) and 2)

percentage lesion volume difference. Differences were compared with related-samples Wilcoxon signed rank tests. Dice similarity coefficient (DSC) was used to compare similarity between automated segmentation methods.

DSC is calculated as:

$$DSC = \frac{2TP}{2TP + FP + FN}$$

Where TP = true positive, FP = false positive and FN = false negative.

Proportion of lesion volume difference between conventional and T2-FLAIR-only segmentation methods was calculated as (T2-FLAIR-only volume – conventional volume / conventional volume). Median percentage volume difference was calculated as median (conventional volume – T2-FLAIR-only volume / average volume)*100.

5.2.8.2 Brain volumetry

Brain volume group means between T1 and T2-FLAIR GIF were compared using paired t-tests. A no-intercept linear regression model was used to compare volume results for three main tissue classes (GM, WM and CSF) and the combined total intracranial volume (TIV) between T1 and T2-FLAIR inputs into the GIF database. The use of a no-intercept model was in line with the expectation of unity between methods and calculations were made for model fit using the Akaike Information Criterion (AIC) for both intercept and no-intercept models.

A subset analysis of the 10% of subjects with the highest WM lesion load was performed in order to assess tissue segmentation performance in the context of more radiologically advanced disease.

The biological utility of T2-FLAIR-only volumetry was assessed by evaluating the ability of the segmentation methods to demonstrate age differences. The CIS cohort that was used for this study had by nature developed little disease-related atrophy, so a general linear model was used to assess brain volume effects of age for both methods. Effect sizes were calculated using Cohen's *f*, where values 0.10, 0.25 and 0.40 represent small, medium and large effect sizes, respectively (Cohen 2013), to determine the number of cases that would

be required to show group differences for age using each segmentation method. Statistical analysis was performed using SPSS Version 25.0.

5.3 Results

This study assessed MRI scans from 66 subjects with CIS. Their mean age (standard deviation) was 34.7 (8.4) and 71% were female. Subjects had a median Expanded Disability Status Scale (EDSS) score of 2.0 (range 0-6.0).

5.3.1 Manual and automated assessment of WMH and MS lesions

Comparison of total lesion volume with Wilcoxon signed rank tests showed statistically significant differences between manual method 1 and all other segmentation methods, manual method 1 producing lower lesion volumes than all other methods at both 1.5T and 3T, $p<0.001$. For manual method 2, at 1.5T, lesion volumes were not significantly different to T2-FLAIR-only segmentation results ($p=0.239$) and at 3T not significantly different to conventional automated (T1+T2-FLAIR) segmentation results ($p=0.231$).

At 1.5T, T2-FLAIR-only lesion volumes were significantly larger than those produced with conventional T1+T2-FLAIR segmentation ($p=0.01$). However at 3T conventional and T2-FLAIR only segmentation results were not significantly different ($p=0.819$).

Median lesion volume in millilitres (ml) and IQR by segmentation method is shown in Table 5-2 and Figure 5-2. An example case showing the segmentation results obtained using the four WM lesion segmentation methods is shown in Figure 5-3.

Table 5-2. Median lesion volume and interquartile range (IQR) for each segmentation method and field strength.

Lesion segmentation method	*Field strength*	*Median lesion volume (ml)*	*Inter-quartile range (IQR)*
Manual 1	1.5T	0.63	2.44
	3T	2.25	3.17
Manual 2	1.5T	3.84	4.83
	3T	5.51	4.88
BaMoS	1.5T	3.38	5.03
	3T	6.48	5.90
T2-FLAIR-only BaMoS	1.5T	4.61	4.81
	3T	6.25	6.95

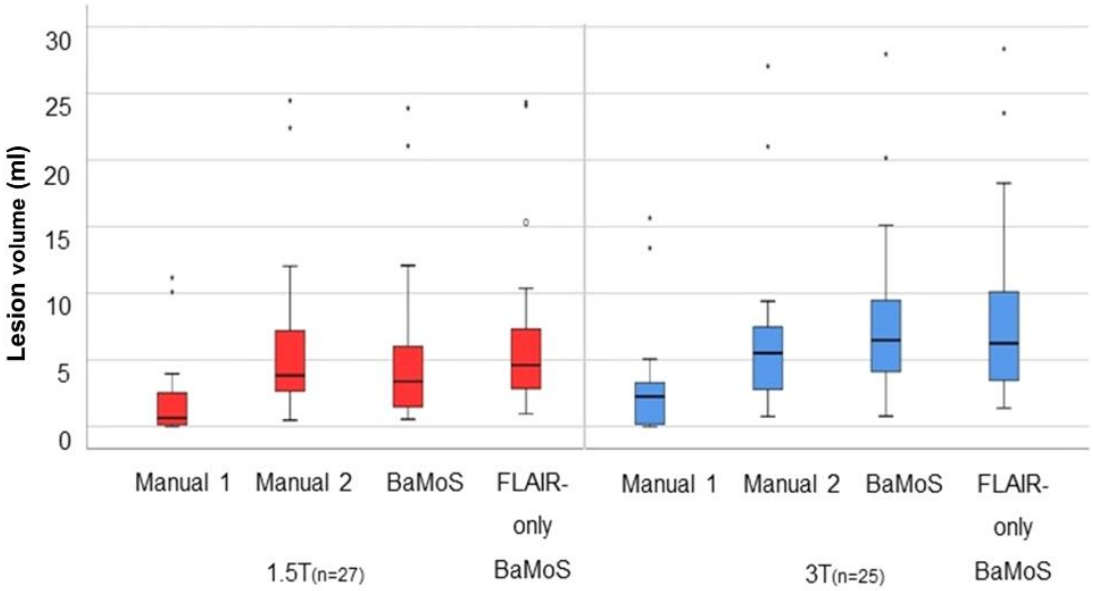


Figure 5-2. Box plots showing lesion volume (median and IQR) by segmentation method and field strength.

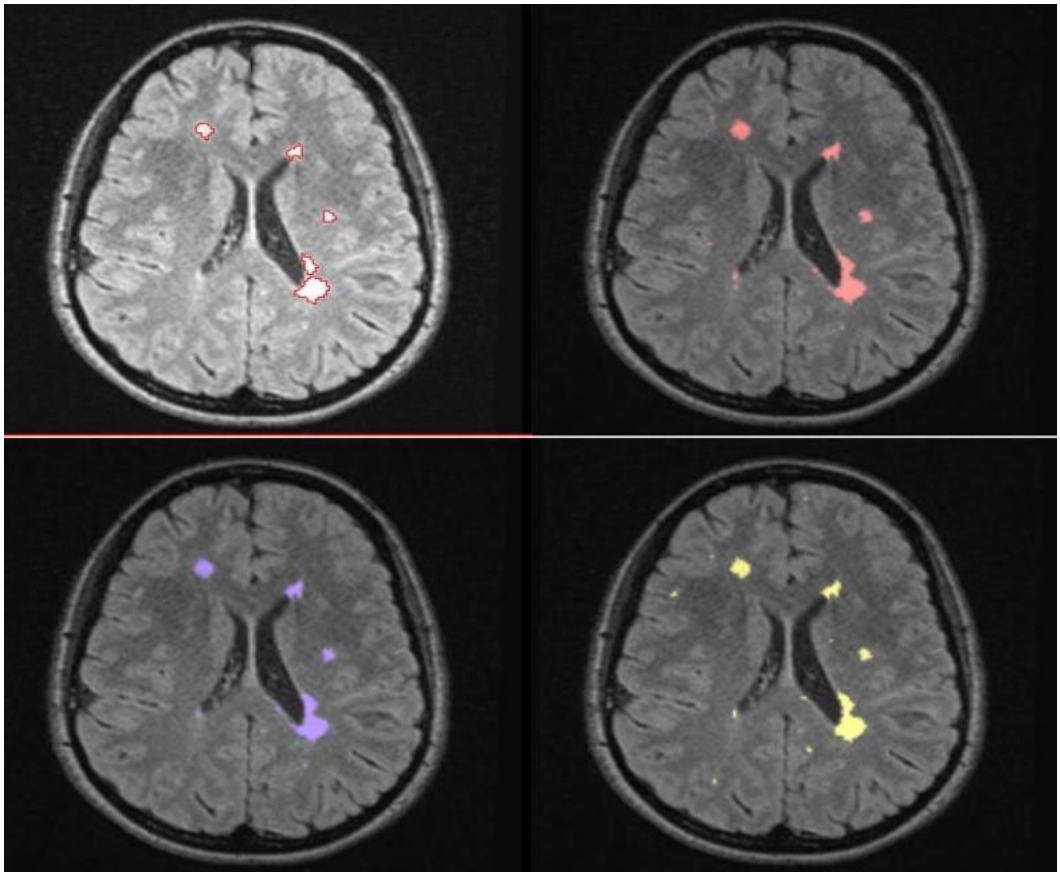


Figure 5-3. An example of WM lesion segmentation results for manual method 1, top left; manual method 2, top right; multi-sequence BaMoS, bottom left; and FLAIR- only BaMoS, bottom right.

Mean DSC (SD) between conventional and T2-FLAIR-only lesion segmentation are 0.46 (0.24) for 1.5T and 0.57 (0.19) for 3T (Figure 5-4). Dice similarity coefficients (DSC) between lesion segmentation methods are shown in Table 5-3.

Table 5-3. Dice similarity coefficients between lesion segmentation methods by field strength. SD, standard deviation.

Lesion segmentation method comparison	*Field strength*	*Dice similarity coefficient Mean (SD)*
Manual 1 vs Manual 2	1.5T	0.21 (0.20)
	3T	0.28 (0.21)
Manual 1 vs BaMoS	1.5T	0.25 (0.23)
	3T	0.32 (0.22)
Manual 2 vs BaMoS	1.5T	0.52 (0.25)
	3T	0.53 (0.24)
Manual 1 vs T2-FLAIR BaMoS	1.5T	0.21(0.21)
	3T	0.29 (0.20)
Manual 2 vs T2-FLAIR BaMoS	1.5T	0.37 (0.23)
	3T	0.43 (0.19)
BaMoS vs T2-FLAIR BaMoS	1.5T	0.46 (0.24)
	3T	0.57 (0.19)

Proportion of lesion volume difference between conventional and T2-FLAIR-only segmentation methods was median (IQR) 0.33 (-1.75 – 1.45) for 1.5T, and -0.13 (-1.87 – 0.18) for 3T. Median percentage volume difference was -28.7% for 1.5T and 13.6% for 3T.

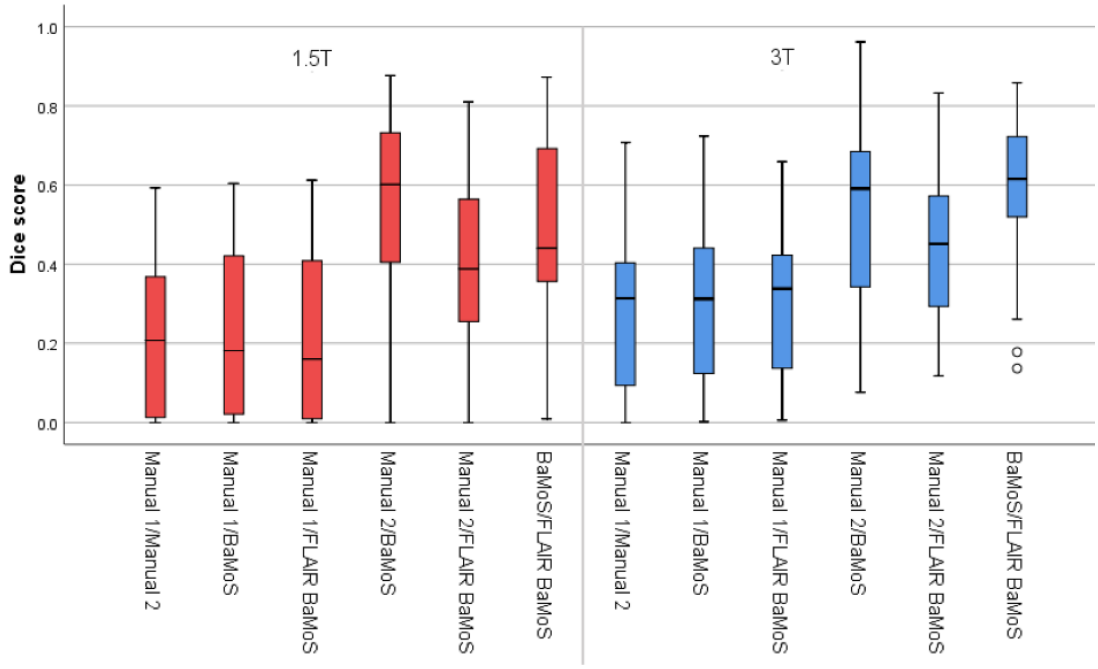


Figure 5-4. Boxplots representing dice similarity coefficient values between methods by field strength.

5.3.2 Brain tissue volumes

Results for each of three key segmentation methods for mean cortical grey matter volume are presented in Table 5-4 by field strength. These are mean volume in ml, (SD) by method 1) original GIF database with T1 input: 503.4 (5.93) at 1.5T and 501.8 (6.10) at 3T, 2) new multi-modal GIF database with T1 input: 515.5 (6.04) at 1.5T and 512.7 (6.12) at 3T, and 3) new multi-modal GIF database with T2-FLAIR input: 529.8 (7.30) at 1.5T and 523.0 (6.77) at 3T.

Results for tissue volume segmentation with varying degrees of WM lesion inpainting show no significant change of GM volume measurements, as shown in Table 5-5. Therefore all results presented here have been processed using WM lesion inpainting from conventional BaMoS WM segmentation. Violin plots for the three tissue classes shown in Figure 5-5 demonstrate that the distribution of tissue segmentation volumes at the individual subject level were very similar for the T1 and T2-FLAIR groups at both field strengths.

Table 5-4. GM volume in ml by GIF method (sequence input and GIF database). Mean volume, standard deviation (SD).

Input	*GIF Database*	*Field strength*	*Mean GM Volume (ml)*	*SD*
T1	original	1.5T	503.4	5.93
	original	3T	501.8	6.10
	new	1.5T	515.5	6.04
	new	3T	512.7	6.12
T2-FLAIR	new	1.5T	529.8	7.30
	new	3T	523.0	6.77

Table 5-5. Mean cortical GM volume as a percentage of TIV, by GIF segmentation method, and by WM lesion inpainting method, for 1.5T and 3T.

CGM segmentation method, 1.5T	*WM lesion inpainting method*	*Mean (SD) GM volume as % of TIV, 1.5T*	*Mean (SD) GM volume as % of TIV, 3T*
T1 (new GIF database)	None	31.64 (1.15)	31.80 (1.10)
	Manual	31.67 (1.15)	31.81 (1.10)
	BaMoS	31.81 (1.22)	31.89 (1.13)
T2-FLAIR (new GIF database)	None	33.20 (1.62)	33.23 (1.43)
	Manual	33.22 (1.63)	33.15 (1.31)
	BaMoS	33.29 (1.66)	33.24 (1.43)

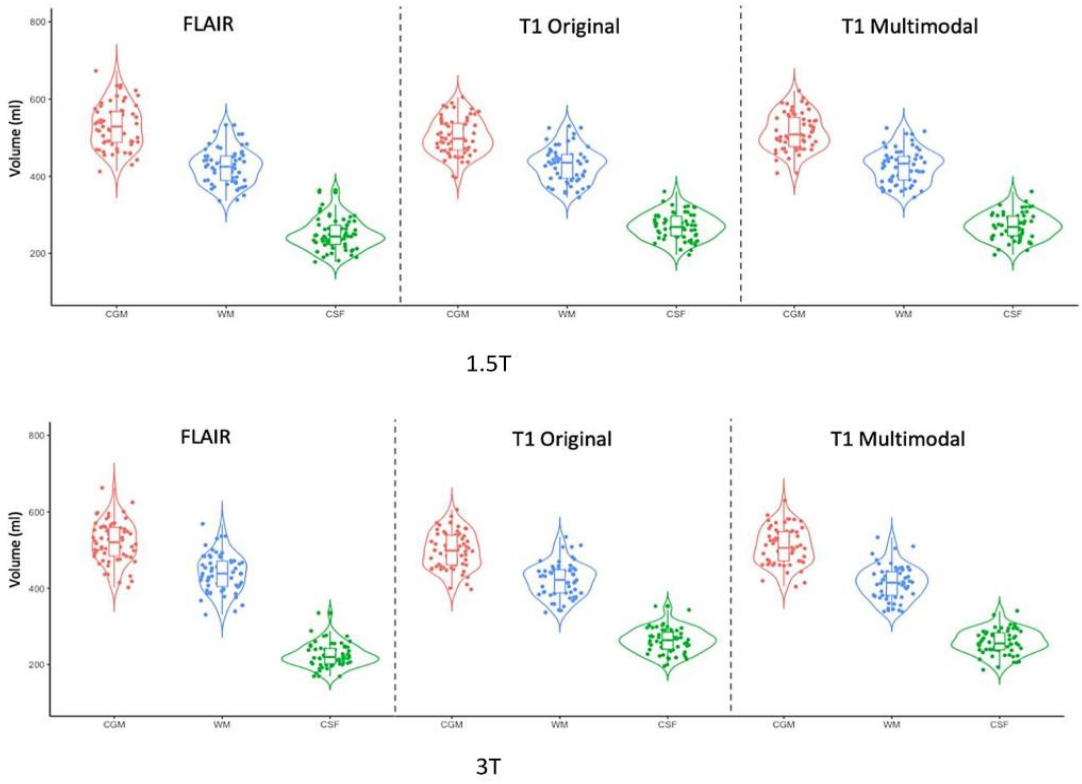


Figure 5-5. Violin plots displaying the actual volumes (in ml) returned per subject by tissue class and field strength – CSF, WM and GM – grouped by segmentation method. FLAIR =adapted GIF database with T2-FLAIR input; T1 original =standard GIF database with T1 input; T1 multimodal =adapted GIF database with T1 input. Violin plots were created using R.

All combinations of paired t-tests performed separately for 1.5T and 3T showed significant differences, $p<0.001$, with T2-FLAIR input producing higher mean GM values at both field strengths. Example subject results are shown in Figure 5-6.

Linear regression modelling for CSF, WM, GM and TIV was performed for these segmentation method comparisons. There was no evidence of model fit deterioration based on AIC calculations (Table 5-6).

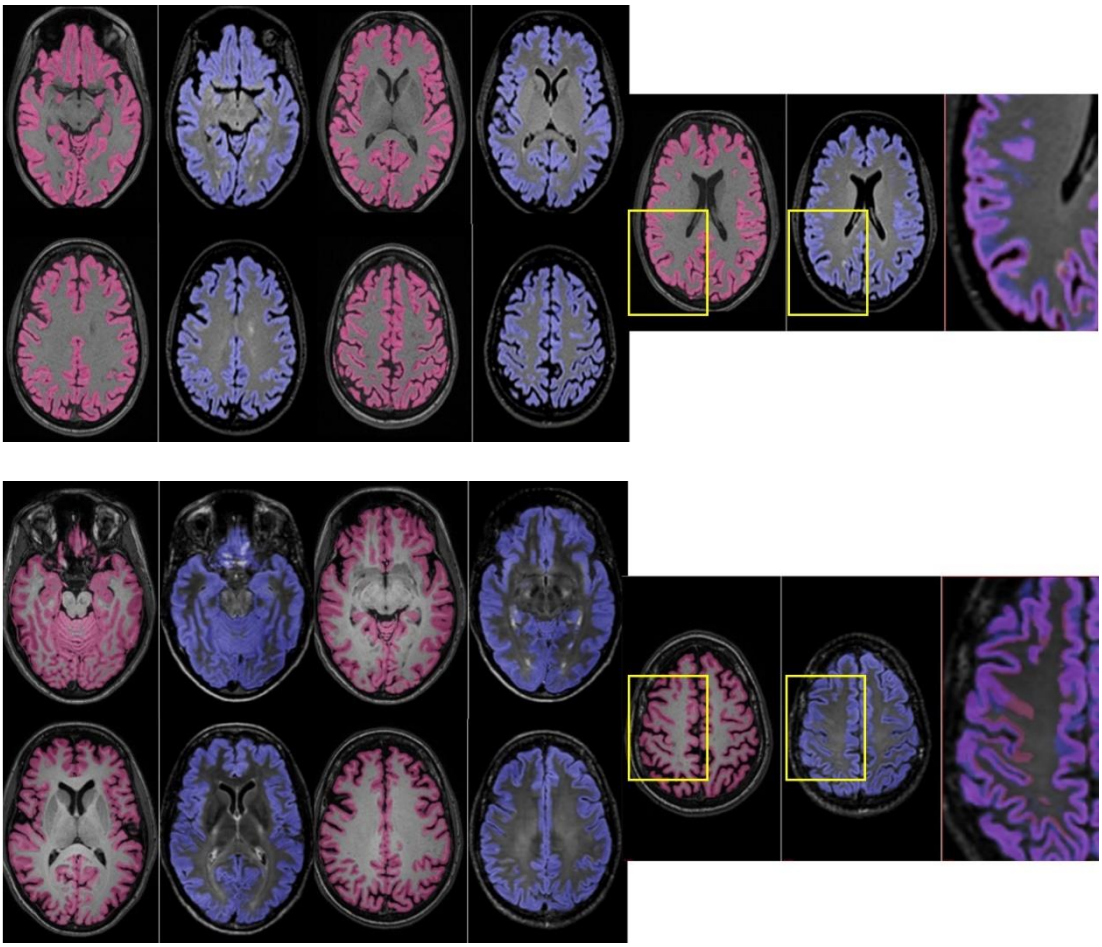


Figure 5-6. A subject's cortical GM segmentation shown for 1.5T (top panel) and for 3T (lower panel), using the multimodal GIF database. T1 segmentation is denoted in pink, and T2-FLAIR segmentation is shown in blue. An enlarged image overlaying both T2-FLAIR and T1 segmentations is included on the right of each series, showing areas of discrepancy, highlighted in the yellow boxes.

Table 5-6. Akaike Information Criterion (AIC) calculations for model fit for both intercept and no-intercept models, for cortical GM volume comparison between GIF methods.

GM volume comparison	*Field strength*	*AIC*	
		Intercept	*No intercept*
T1 (original vs. new database)	1.5T	234.63	233.55
	3T	220.83	221.78
T1 vs T2-FLAIR (new database)	1.5T	432.00	430.00
	3T	397.06	395.23

Results of T1 segmentation using the original and new GIF databases are presented in Table 5-7. For T1 and T2-FLAIR results using the new GIF

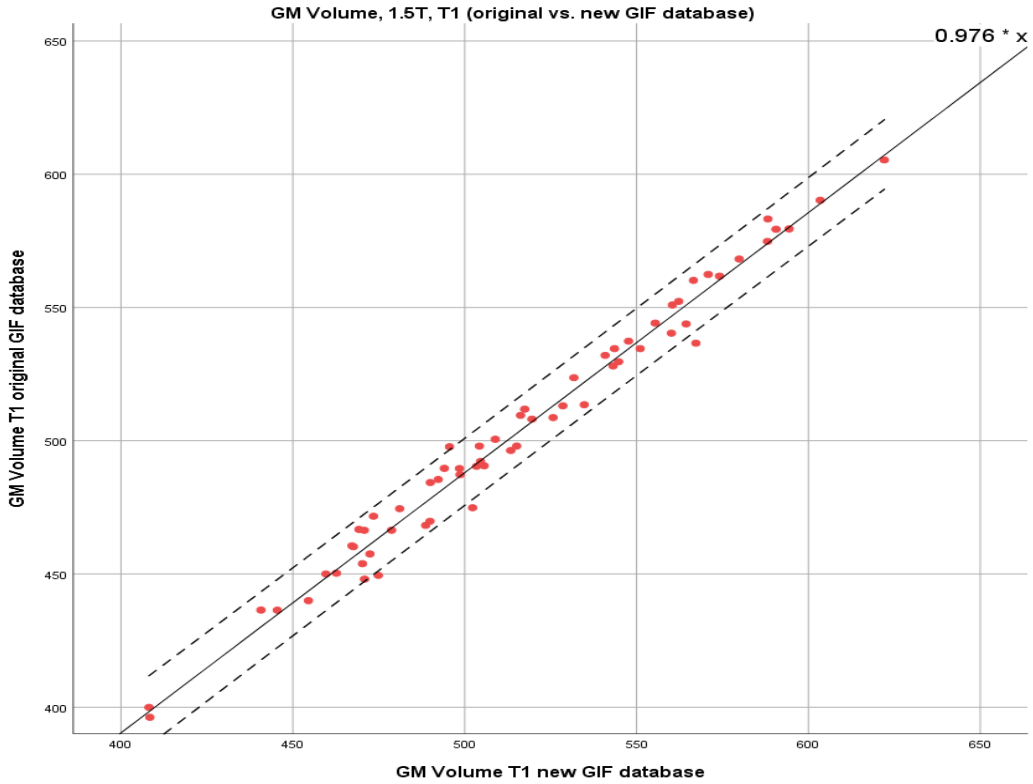
database, linear regression model results are shown in Table 5-8. For GM volume using the new database, the model showed at 1.5T that R^2 was 0.997, β (SE) 1.028 (0.007), and at 3T that R^2 was 0.998, β (SE) 1.019 (0.006). Figure 5-7 and Figure 5-8 show the GM correlations for change in GIF database and change of input sequence by field strength. These demonstrate a widening of the 95% confidence intervals for the correlations between GM volumes derived using T1 and T2-FLAIR input.

Table 5-7. Linear regression outputs for comparison of T1 inputs into the original T1-only and new GIF database using a no-intercept model. β=slope coefficient, SE=standard error

Tissue	*Field strength*	*β (SE)*	*R^2*
GM	1.5T	1.054 (0.008)	0.996
	3T	1.041 (0.006)	0.998
WM	1.5T	0.987 (0.008)	0.996
	3T	1.036 (0.009)	0.995
CSF	1.5T	0.913 (0.012)	0.988
	3T	0.850 (0.008)	0.994
TIV	1.5T	0.984 (0.004)	0.999
	3T	0.988 (0.004)	0.999

Table 5-8. Linear regression outputs for comparison of T1 and T2-FLAIR inputs into the new GIF database. β=slope coefficient, SE=standard error

Tissue	*Field strength*	*β (SE)*	*R^2*
GM	1.5T	1.028 (0.007)	0.997
	3T	1.019 (0.006)	0.998
WM	1.5T	0.995 (0.007)	0.997
	3T	1.055 (0.008)	0.996
CSF	1.5T	0.944 (0.012)	0.989
	3T	0.859 (0.009)	0.994
TIV	1.5T	0.973 (0.004)	0.999
	3T	0.999 (0.004)	0.999



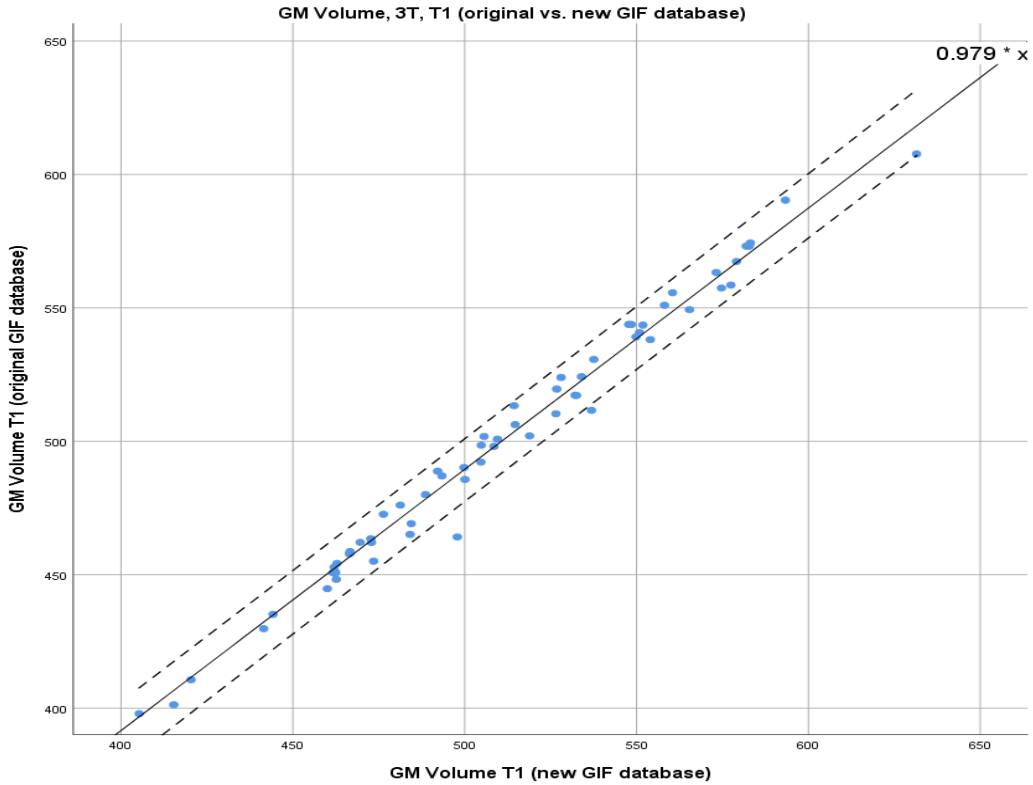


Figure 5-7. Scatter plots for GM volumes in ml; T1 input into conventional and new GIF database. Top graph: 1.5T. Lower graph: 3T. Coefficient shown in upper right-hand corner and 95% CI shown with dotted lines.

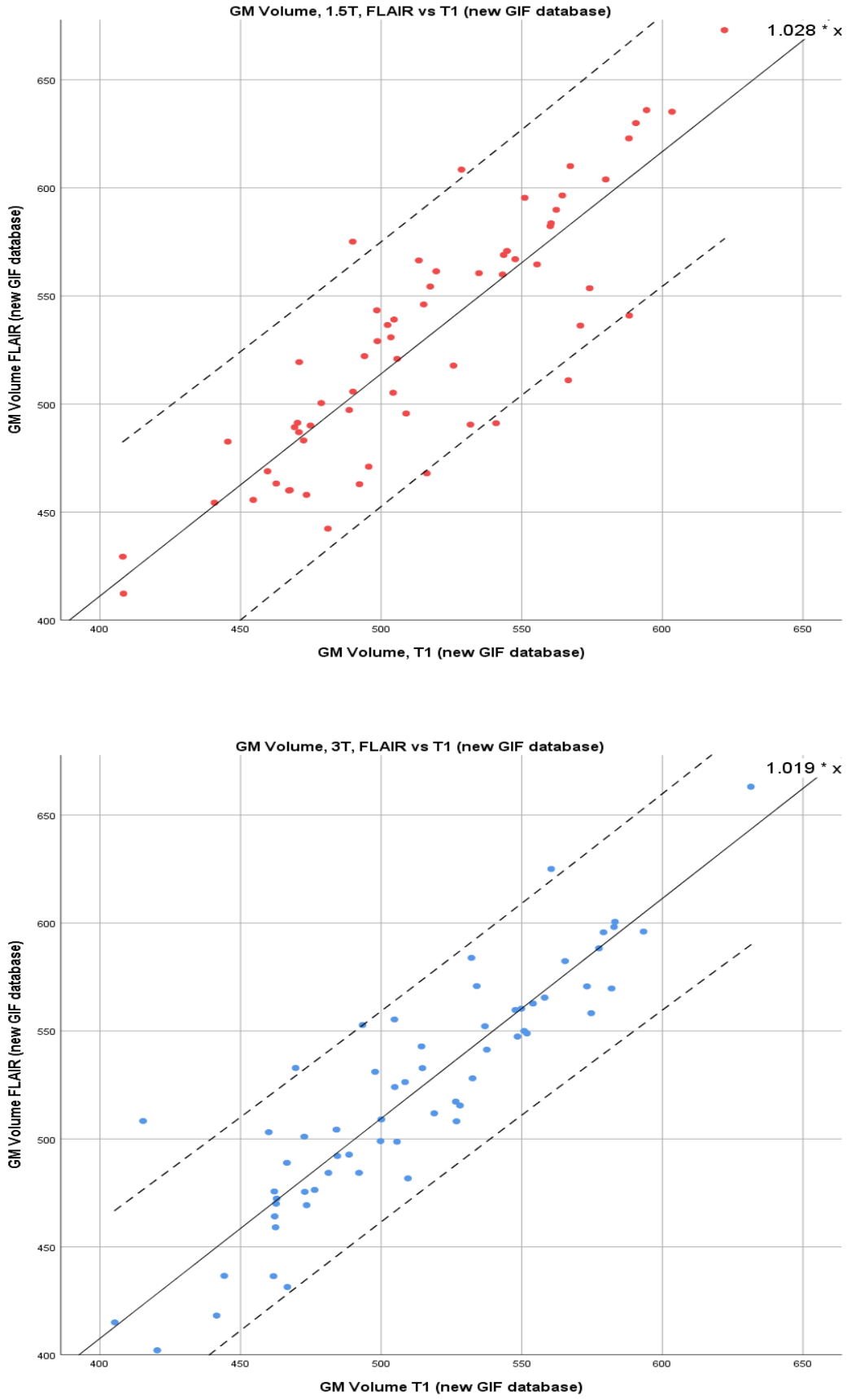


Figure 5-8. Scatter plots for GM volumes in ml; T2-FLAIR vs. T1 input into new GIF database. Top panel: 1.5T. Lower panel: 3T. Coefficient shown in upper right-hand corner and 95% CI shown with dotted lines.

A subset analysis of tissue segmentation results was performed for the 10% of subjects with the highest lesion loads, in order to address generalisability of these results to other MS populations. In this subset the lesion volume in millilitres as calculated using conventional BaMoS was mean (SD) 14.1 (5.8) at 1.5T and 15.5 (6.5) at 3T. Linear regression results for GM segmentation between T1 and T2-FLAIR for this subset using the new GIF database were β (SE) 1.029 (0.024) and R^2 0.997 for 1.5T and 1.022 (0.019), R^2 0.998 for 3T, as shown in table X. An example of GM segmentation results for a subject with a high WM lesion load is shown in Figure 5-9.

For each segmentation methods univariate analyses were computed for GM volume versus age, which showed that GM volumes were significantly associated with TIV and age and therefore these were included as covariates for all subsequent models. Field strength was included as a fixed factor. For all three of the GIF database sequences and input combinations, age was a significant covariate: 1) conventional T1 GIF (R^2=0.999, standard error (SE)=0.178 p=0.001), 2) T1 using the new GIF database (R^2=0.999, SE=0.182, p<0.001), and 3) T2-FLAIR using the new GIF database (R^2=0.998, SE=0.247, p=0.005). Effect sizes for age, Cohen's *f*, were for T1 GIF *f*=0.36 and for T2-FLAIR GIF *f*=0.27.

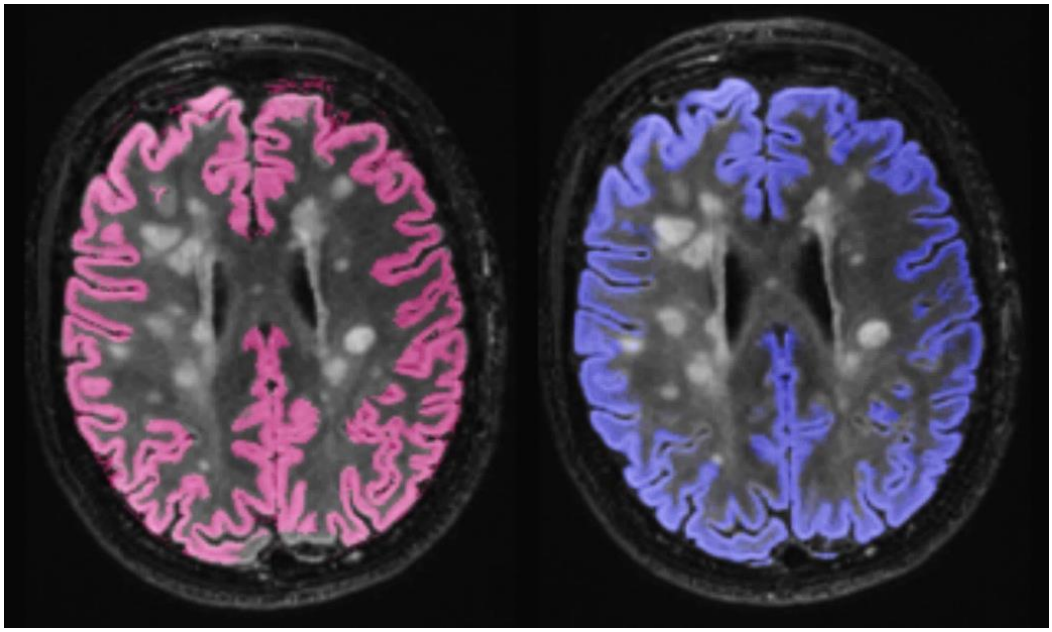


Figure 5-9. GM segmentation performance in the context of high WM lesion load, using the new GIF database (pink = T1, blue = FLAIR).

5.4 Discussion

This study aimed to investigate how automated T2-FLAIR-only lesion and brain segmentation would compare to conventional segmentation methods in a group of CIS subjects at two field strengths. Clinical MS imaging protocols often do not include a volumetric T1 sequence, which is usually required by standard T1 or multi-sequence automated quantification techniques, thereby limiting the use of these methods in clinical settings.

Therefore, this study hypothesised that results of T2-FLAIR-only segmentation would provide results that are comparable to T1 and multi-sequence methods. A cohort of CIS subjects from multiple centres who had been scanned at 1.5T and 3T in the same week was used to compare the output of WM lesion and brain tissue volume segmentation using established methods, namely BaMoS and GIF algorithms, to that from adapted T2-FLAIR versions of these tools.

Lesion segmentation with the automated T2-FLAIR-only method was comparable to conventional automated segmentation at 3T. At both 1.5T and 3T, brain tissue segmentation was robust using the T2-FLAIR method, as evidenced by high R^2 linear regression values and maintained age-related brain volume change discrimination.

5.4.1 WM lesion segmentation

Two types of manual WM lesion segmentation were used for comparison of the automated methods, one based on expert consensus reading of MS-specific lesions and the other that included all WM hyperintensities in the image and not only MS-specific lesions, at 1.5T and 3T. The WM hyperintensities that were not identified by experts to be MS lesions could include non-specific WM lesions consistent with vascular disease or aging, as well as periventricular caps and bands, or indeed image artefacts. There may also be some which are true MS lesions that are not captured by conservative inclusion criteria.

Results of the two manual segmentation methods varied from each other quite considerably, and automated segmentation results were much more similar to the second manual method. This finding highlights that classical segmentation

algorithms can be limited in their ability to discriminate true MS-type lesions from any WM hyperintensity.

It is important to consider that this may be an inherent limitation in applying intensity-based methods to quantify MS-specific pathology, and may rely on visual quality control measures or ensuring that eventual clinical end-users are aware of the limitation if full automation is implemented.

There were differences in lesion segmentation performance between field strengths, as discussed further in section 5.4.3. Automated T2-FLAIR-only lesion segmentation results at 1.5T were comparable to a manual method which segmented all WM hyperintensities (manual method 2). At 3T lesion volumes were not significantly different between conventional and automated methods, and proportional lesion volume differences were very small. This contrasted with the scenario at 1.5T, where lesion volumes were not comparable between the two automated methods.

Use of a CIS cohort meant that relatively lower WM lesion loads were expected, which in turn made lesion segmentation method comparison challenging and produced low dice scores. It is accepted that accurate automated lesion segmentation is more accurate at higher lesion loads (Carass et al. 2017). An important future step would be to apply this method to an MS population with higher lesion loads.

5.4.2 Brain tissue segmentation

The T2-FLAIR-only brain tissue segmentation method tested here generated similar results to the conventional T1 method, with very high R^2 and low standard error values. The coefficients quoted in tables 6 and 7 can be interpreted as simple multiplicative factors between the two methods and their raw sizes demonstrate minimal differences in brain tissue volume measurements between change of GIF database, sequence input, and a combination of both changes. In a subset analysis of cases with high WM lesion volumes, robust tissue segmentation performance was maintained.

T2-FLAIR-only GIF also maintained the demonstration of biological effects in the study population, as subject age remained a highly significant association

with GM tissue volume. The age-related effect sizes were of a similar magnitude between T1 and T2-FLAIR GIF.

These results support the potential utility of T2-FLAIR-only automated brain tissue segmentation as a clinical tool for brain volume analysis. Further work would be needed to establish its validity in other MS phenotypes where there is likely to be more marked parenchymal atrophy present. Currently it is not an expected element of standard clinical care to report the neurodegenerative aspect of MS. Even though it is recognised as an important biomarker in the research setting, it faces barriers to clinical adoption (Sastre-Garriga et al. 2020). Automated segmentation tools could assist in this translational challenge (Sormani et al. 2017), however technical hurdles such as the one explored here need to be addressed. Many clinical centres still use a non-volumetric T2-FLAIR sequence in their MS protocols, and there are useful tools that have been shown to be able to measure central atrophy accurately from heterogeneous 2D T2-FLAIR data (Zivadinov et al. 2018). However there is increasing adoption of a 3D T2-FLAIR sequence by clinical centres for their MS imaging in line with the most current guidance (Filippi et al. 2019; Saslow et al. 2020; Sastre-Garriga et al. 2020), in parallel with an increasing clinical interest in adoption of quantification techniques, making this work timely and relevant.

5.4.3 Field strength and acquisition

This study has shown that T2-FLAIR-only tissue segmentation is robust, with the application of small multiplicative differences between volumes obtained with T1-based segmentation. There is some variation in performance between field strengths, with differences in the multiplicative factors and slightly lower variance at 3T than 1.5T. As already discussed, WM lesion volume results were overestimated at 1.5T. These points should be considered when planning to implement automated segmentation methods for clinical use; results for different patients and at different timepoints may not be directly comparable if there is a change in field strength (Han et al. 2006; Lysandropoulos et al. 2016).

Technical differences between scanners and acquisition parameters can

impact on the performance of automated segmentation algorithms, even within a single field strength (Biberacher et al. 2016), which is a fundamental issue to consider in the clinical setting. Automated segmentation methods utilising T1 images have been shown to be sensitive to differences in sequence parameters on the same scanner amounting to volumetric errors of 4-5% at 1.5T, which would obscure biological effects (Haller et al. 2016). This issue is compounded by the limited experience to date in standardising T2-FLAIR acquisition protocols, in contrast to the advances that have recently been seen with T1 imaging standardisation (George et al. 2019; Jack et al. 2015). Initiatives like ADNI have been leading the efforts to standardise protocols and minimise these sources of bias (Brewer 2009).

Similar work towards standardisation of 3D T2-FLAIR acquisition may address a significant amount of variability, at least across a single clinical service initially and ultimately across centres to facilitate research and data sharing. Harmonisation initiatives for T2-FLAIR with direct applicability to MS imaging are being championed by MAGNIMS, the North American Imaging in MS Cooperative (www.naimscooperative.org, NAIMS 2021) and the Consortium of Multiple Sclerosis Centres (www.mscare.org, CMSC 2021), (Saslow et al. 2020; Wattjes et al. 2021). Adoption of these efforts will facilitate validation and interpretation of results of automated segmentation algorithms in the clinical setting.

5.4.4 Limitations

Several limitations affected this study. Whilst the dataset was multi-centre and multi-vendor, therefore providing a realistic mimic of a clinical dataset, subject numbers per centre were not balanced and image homogeneity was not guaranteed. Use of a CIS cohort meant that it was difficult to address the effect of disease-mediated brain atrophy on T2-FLAIR-only brain tissue segmentation, which is important for MS imaging. A subset analysis of CIS subjects from the cohort with high lesion loads did however show consistently good results for segmentation performance. Additionally, scan-rescan reproducibility within each field strength was not tested for brain segmentation measurements due to the lack of an available dataset.

5.4.5 Conclusions

This study has shown that T2-FLAIR-only automated brain volume segmentation is comparable to conventional T1 or dual-modality methods, with lower lesion segmentation reliability at lower field strengths. Future work with other MS phenotypes, combined with efforts towards clinical image acquisition harmonisation, would further improve clinical validation. This important translational task is one aspect of a wider challenge. For the provision of fully automated, robust quantification methods for clinical use, ongoing efforts need to be pursued in terms of standardisation of imaging protocols and validation of quantification methods at the individual patient level. These issues must be addressed if WM lesion and brain volume analysis is to be widely adopted and meaningfully utilised in radiological MS reporting for patient benefit.

5.5 Towards a reporting tool for clinical MS application

5.5.1 A FLAIR-based MS reporting tool

Based on the previous work, I have been working towards the design and construction of a quantitative report based on data extracted from FLAIR images for use in clinical MS reporting.

The aim is to summarise and present the quantitative results for an individual MS patient contextualised with normative and MS population reference data. The reporting tool would aim to assist the radiologist's routine reporting workflow. It would focus on the analysis of demyelinating lesions, and supplement this with pertinent regional brain volume information.

Useful information that could be extracted would include volume, number and regional distribution of white matter lesions. Lesion load could be presented within the four diagnostic regions (periventricular, juxtacortical, deep white matter, and infratentorial) that define dissemination in space. Providing regional lesion volume data may be more useful than a single summated WM lesion volume which may be difficult to interpret longitudinally, given that lesions may shrink as well as grow or appear de novo between timepoints (Dwyer et al. 2018). In addition, the individual's WM lesion load could be contextualised with reference data taken from a healthy control population and a group of subjects with MS.

Given the challenges to accurate interpretation of summated WM lesion volume measurements, quantification of brain volumes may arguably be more useful. These could be presented in reference to the expected age-matched normative reference and the MS population would provide an indication of the severity of neurodegeneration affecting an individual patient. However providing generalisable reference ranges in an MS population of mixed disease durations is challenging and likely to be inherently uncertain. In a similar vein to the quantitative report for dementia discussed in chapter 4, a graph displaying global brain volume as measured by brain parenchymal fraction (BPF) plotted on the reference data, as well as a 'rose plot' of focal regions that are known to be particularly affected by MS – for example the thalamus, precuneus, hippocampus, and brainstem – may be included.

There is an interesting opportunity to include an additional QIB that has been shown to be useful in an MS population – namely predicted brain age (Cole et al. 2020). The brain age model uses machine learning to predict chronological age from MRI data (Cole and Franke 2017). It can be applied to identify individuals with a gap between their chronological age and the accelerated ageing of their brain as a potentially important imaging biomarker which can predict mortality (Cole et al. 2018). It has previously been used to demonstrate a 'brain age gap' between chronological age and accelerated brain ageing due to a range of pathologies, including dementia (Franke and Gaser 2012), schizophrenia (Koutsouleris et al. 2014) and traumatic brain injury (Cole, Leech, and Sharp 2015). When it was applied to an MS patient cohort it was sensitive to disease-related brain atrophy and correlated with clinical progression (Cole et al. 2019, 2020). Further studies are needed to establish whether a cross-sectional brain age metric is useful for prognostication in MS. To date, the technique has not yet been fully applied to volumetric FLAIR images. Inclusion of this relatively new QIB that has been validated in the research setting into a reporting tool, following technical validation with FLAIR input, would allow for its technical and clinical validation with clinical grade data.

5.5.2 Methods

5.5.2.1 Reference data

In order to work towards the development of an MS reporting tool, appropriate FLAIR reference data from healthy control and MS populations is being established having been granted UCL ethical approval.

Several open-source datasets have been identified which include a 3D FLAIR sequence in their protocols. However, the majority of these datasets contain older subjects above the age of 50, whereas MS is a disease that commonly begins in younger adults. Therefore, the MAGNIMS group was invited to contribute data for QNI and FLAIR brain age validation. This MAGNIMS collaboration has meant that MS centres from across Europe have shared 3D T1 and FLAIR data from healthy control populations covering an expanded age range, from 18 to 93 years. The data sources are now explained.

EPAD is the European Prevention of Alzheimer's Disease Consortium (ep-ad.org, EPAD 2021), which is conducting multi-site pan-European imaging and cognitive and biomarker research (Solomon et al. 2018). Those EPAD subjects with an overall Clinical Dementia Rating (CDR) score of zero were selected to be included in this healthy control reference population.

ADNI-3 is the latest cohort of the Alzheimer's Disease Neuroimaging Initiative (adni.loni.usc.edu/adni-3/, ADNI 2021). Those subjects defined as healthy controls by the study were included. NIFD is the common name given to the University of California San Francisco (UCSF) frontotemporal lobar degeneration neuroimaging initiative (memory.ucsf.edu/research/studies/nifd, UCSF 2021). The control group from this study was used here. MPI refers to the Max Planck Institute Leipzig Mind-Brain-Body dataset (Mendes et al. 2019).

The following MAGNIMS centres have contributed data so far: Vall d'Hebron University Hospital, Barcelona, Spain; the San Raffaele Scientifica Insititute, Milan, Italy; University of Siena, Siena, Italy; and the Medical University Graz, Graz, Austria. More data is anticipated from MAGNIMS centres in Oxford, Naples, Hanover, Oslo and Mainz.

5.5.2.2 Processing pipeline

Each subject's 3D FLAIR scan is processed using the following steps. Firstly N3 bias field correction is performed to correct for intensity nonuniformity (Sled, Zijdenbos, and Evans 1998). Next, lesion segmentation is performed using the nicMS algorithm (Valverde et al. 2017), which is a freely available patch-based convolutional neural network (CNN) method (github.com/sergivalverde/nicMSlesions/, Valverde 2021). Lesion segmentation results are then used to perform lesion filling (Prados et al. 2016). Brain tissue segmentation with GIF is then performed using the lesion-filled image (Cardoso et al. 2015). BPF is computed from the GIF outputs. Finally, the WM is split into layers which define lesion location (Sudre et al. 2018), and each lesion is assigned to one of the following location definitions: cortical GM, juxtacortical, deep WM, periventricular, cerebellum, brainstem, deep GM, leukocortical, or mixed WM/GM.

5.5.3 Preliminary results

The normative reference dataset characteristics are described in Table 5-9. Graphs in Figure 5-11 show the age distribution of this preliminary healthy control reference dataset followed by scatter plots of lesion count, lesion volume, BPF, cortical GM and WM volumes. Example lesion segmentation results from the processing pipeline are shown in Figure 5-10.

Table 5-9. Summary characteristics of each component of the healthy control reference dataset. M=male, F= Female. SD=standard deviation.

Data source	*Number of subjects*	*M:F*	*Age Mean (SD)*	*Field Strength*
EPAD	832	347:485	64.2 (6.8)	1.5T and 3T
ADNI-3	99	48:51	76.5 (8.0)	3T
NIFD	116	51:65	62.9 (7.6)	3T
MPI	86	73:13	34.4 (17.7)	3T
Barcelona	46	13:33	44.5 (9.7)	3T
Milan	186	100:86	33.5 (11.9)	3T
Siena	27	15:12	36 (8.4)	3T
Graz	33	13:20	30.5 (8.9)	3T
Total	1425	660:765	57.2 (16.5)	

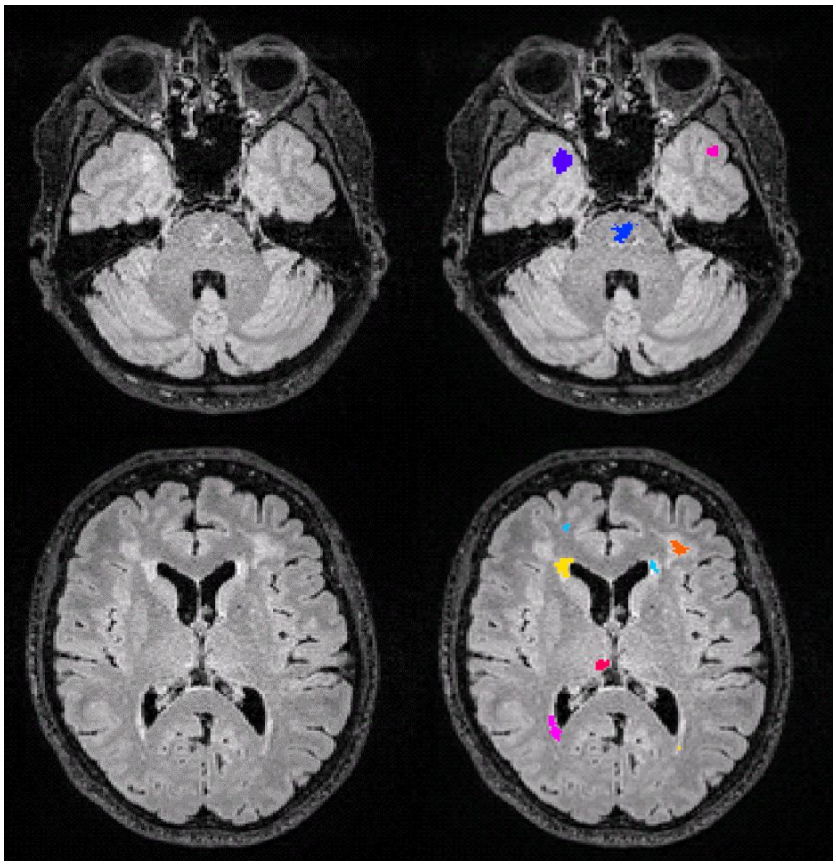


Figure 5-10. Example lesion segmentation results for a healthy control subject from the ADNI cohort of the normative reference database.

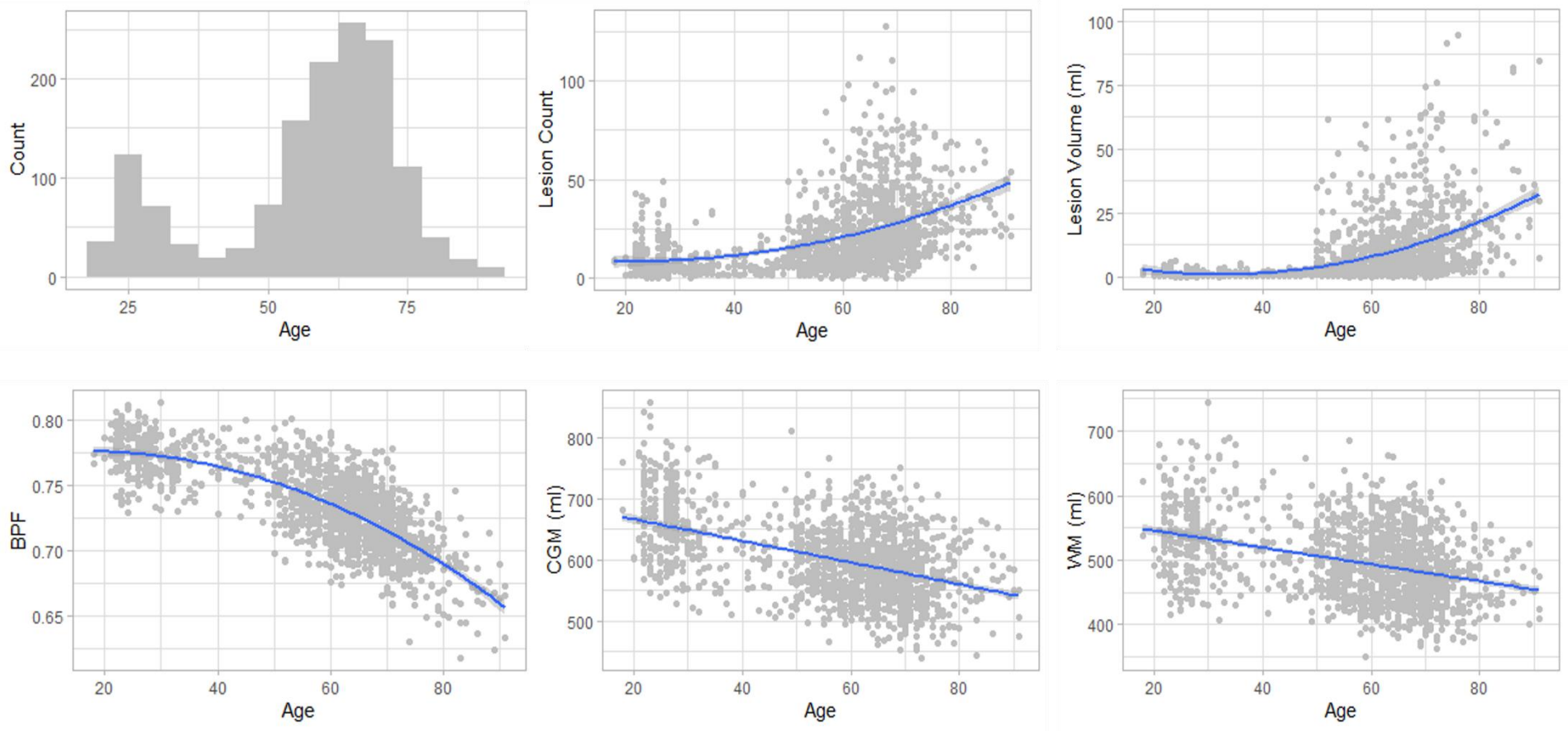


Figure 5-11. Results for the healthy control reference dataset. BPF = brain parenchymal fraction. CGM = cortical grey matter, WM = white matter, ml=millilitres.

5.5.4 Ongoing work towards an MS reporting tool

Establishing reference data forms the basis of the MS quantitative report. The strengths of this healthy control reference dataset include its multi-centre nature and good coverage of the age range of 18-93 years. Ongoing work is focusing on establishing a reference MS population dataset that will be processed using the same segmentation pipeline.

The aim will be to display the healthy control and MS reference curves on a single graph, showing the differences between brain and lesion volumes in healthy versus subjects with MS. A gap between the two curves for brain volumes would reflect the neurodegenerative element of MS pathology (as depicted in Figure 5-12). The individual subject whose images are processed using the quantitative reporting tool would have their measurements plotted in the context of both of these reference populations.

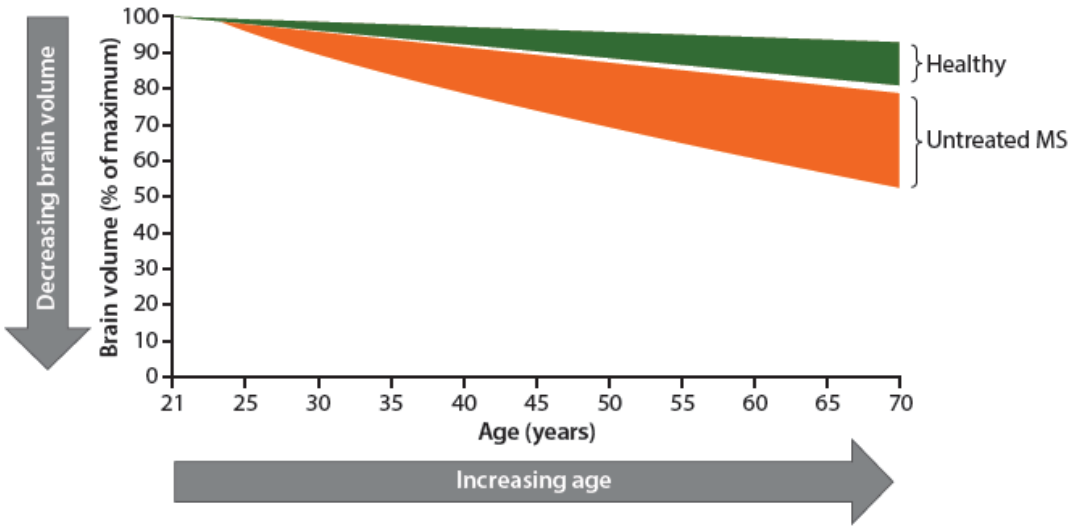


Figure 5-12. Difference in brain volume with age between healthy and MS populations. Figure from Giovannoni et al. 2016.

A mock-up version of the planned final MS quantitative reporting tool is shown in Figure 5-13. Patient and scan details would be followed by snapshot segmentation results overlaid on axial slices of the subject's FLAIR scan. Results for lesion load and lesion count per diagnostic region would be tabulated alongside normalised brain volume and the subject's predicted brain age.

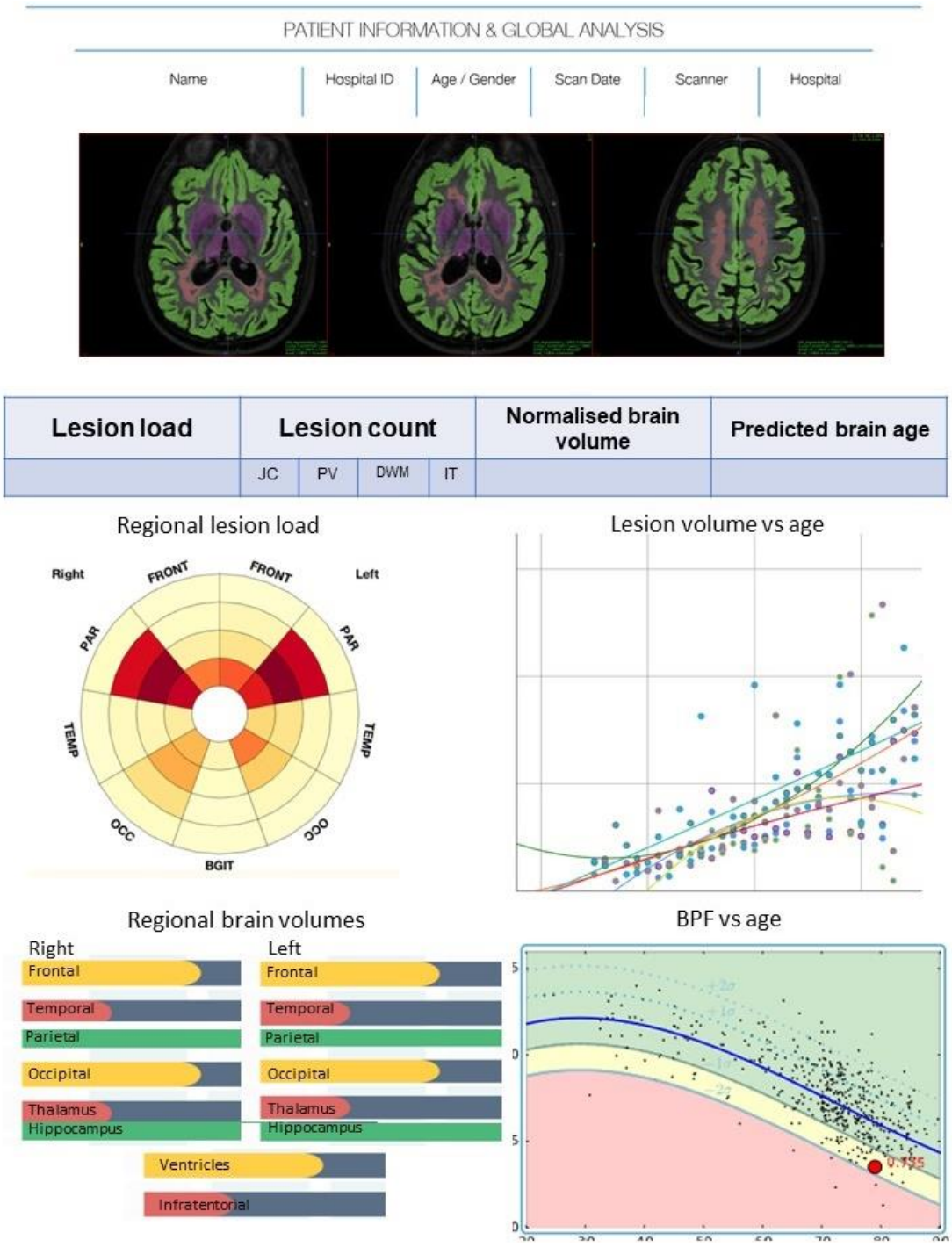


Figure 5-13. A plan for the content and layout of a quantitative report for MS.

Lesion load per region could then be represented using a bullseye plot, accompanied by a scatter plot of lesion volume by age which compares healthy and MS reference populations. Regional brain volumes that are particularly relevant in MS could be presented using a traffic light system referenced against the healthy control population. Finally, BPF could be displayed on a

reference graph which presents an MS and healthy control population plotted simultaneously.

Performing pre-use clinical validation in the form of credibility and accuracy studies, using similar methods to those demonstrated in previous chapters, will be fundamental for assessing the clinical utility and potential benefit of implementing this MS quantitative report. Future work that could potentially build on this initial report construction would be to work towards a longitudinal assessment tool which could quantify changes in brain and lesion volume over time and calculate the likely disease trajectory. There are also potentially useful extensions to the brain age concept, for example modelling predicted EDSS based on brain imaging appearances, which would require rigorous validation.

6. Structured reporting for gliomas based on VASARI criteria

6.1 Introduction

6.1.1 Gliomas

Gliomas are a diverse group of tumours that arise from the glial cells of the central nervous system. They are the commonest intrinsic primary brain tumours and account for the majority of malignant brain tumours (Ostrom et al. 2017). Gliomas arise from various glial cell types and are classified into four World Health Organisation (WHO) grades which reflect their pathological and molecular genetic features (Louis et al. 2016). Grades I and II are also referred to as low-grade gliomas (LGG) and grades III and IV as high-grade gliomas (HGG). Higher grading is associated with more aggressive features and lower median survival. While LGGs have lower mortality, they are associated with morbidity including epilepsy and impaired cognitive function, and there is a high recurrence rate after LGG resection (Buchlak et al. 2021).

Molecular genetic features appear to be the most important in determining the aggressiveness of glioma subtypes, which can differ from their histological subtypes (Thust et al. 2018). The WHO classification includes three main categories of adult diffuse glioma based on isocitrate dehydrogenase (IDH) gene mutation and chromosomal co-deletion of chromosome arms 1 and 19 (1p/19q co-deletion). Gliomas of equivalent cell type and grade that are IDH mutant have a more favourable prognosis compared to those that are IDH-wild-type (Suzuki et al. 2015). Definitive glioma grading requires tissue for histology, the results of which determines ongoing treatment strategy. The mainstay of treatment is maximum safe surgical resection and radiotherapy often combined with adjuvant temozolomide chemotherapy (Stupp et al. 2005).

6.1.2 Clinical MRI for gliomas

Features on structural MRI can help neuroradiologists to predict glioma type and severity, as well as being fundamental to surgical planning and monitoring of treatment response. Certain important structural MRI features are more suggestive of low- or high-grade pathology. If a glioma appears low-grade it may be managed by interval imaging assessment as opposed to urgent neurosurgical intervention. Whilst structural MRI is not used in isolation when

making these important assessments, objective agreement and clear reporting of structural MRI features facilitates diagnosis and the subsequent decision-making process for the multidisciplinary team and the patient.

The use of multi-sequence MRI protocols allow various glioma features to be highlighted according to underlying tissue characteristics, including peri-tumoural oedema / infiltration, tumour necrosis and haemorrhage. However, it should be noted that conventional MRI signal properties cannot distinguish, for example, between peri-tumoural oedema from tumour invasion. A typical protocol would include pre- and post-contrast enhanced T1-, T2- and FLAIR-weighted sequences as well as diffusion weighted imaging (DWI). This protocol is recommended as the minimum for clinical use by the European Society of Neuroradiology (ESNR) (Thust et al. 2018), based on clinical trial standards set out by the European Organisation for Research and Treatment of Cancer (EORTC) and the United States National Brain Tumour Society (NBTS) in the EORTC-NBTS protocol (Ellingson et al. 2015).

Structural and DWI sequences are often supplemented with advanced imaging to further explore specific characteristics, for example perfusion weighted MR imaging and magnetic resonance spectroscopy, which can further elucidate tumour subtype or increase the suspicion of high grade features (Fouke et al. 2015). These advanced imaging techniques are used by approximately two-thirds of European imaging departments, as reported in a recent survey of 220 centres (Thust et al. 2018).

The heterogeneity of gliomas is reflected by the wide variety of associated imaging features. Much research has focused on which MRI features are associated with particular histological and molecular genetic features, with subsequent management and prognostic relevance. Prediction of glioma genotype from imaging phenotypes – referred to as radiogenomics – is developing apace (Smits 2021). The preoperative MRI study is key for neurosurgical planning, by informing whether the patient should undergo biopsy or resection, delineating safe resection margins, and highlighting the potential risk of complications and involvement of eloquent brain functions.

Tumour size and location features additionally inform oncological management by determining suitability for radiotherapy.

6.1.3 Aims of this chapter

Current clinical neuroradiology reporting practice relies on the individual style of the neuroradiologist, informed by their own training and experience. While certain imaging features may be frequently mentioned in reports for gliomas others may be less reliably included. In particular, communication of quantitative information like lesion size or proportion of enhancement may benefit from standardisation for diagnostic characterisation as well as monitoring any change in the tumour appearance on serial imaging. Work towards structured reporting of brain tumours in the clinical setting has shown increased reliability of feature detection compared to free-text reporting (Bink et al. 2018; Zhang et al. 2019). To this end, the American Society of Neuroradiology, American College of Radiology and the Radiological Society of North America (RSNA) have collaborated on a Common Data Elements (CDE) project, which sets out to achieve agreed uniform essential concepts or features that should be included in a radiology report for a given clinical indication.

The Visually Accessible Rembrandt Images feature-set or VASARI criteria (wiki.cancerimagingarchive.net) are a set of 26 standardised imaging features which describe characteristics of gliomas on structural MRI. They have been validated based on the glioma literature as the most useful set of imaging features from a large dataset of baseline HGG and LGG imaging studies, and were shown to correlate with tumour genotype on pathological assessment (Gutman et al. 2013). They have been shown to be effective in predicting treatment outcome and survival in the context of both low and high grade gliomas (Wangaryattawanich et al. 2015; Zhou et al. 2017).

The CDE project suggests as a key example that a set of CDEs could be directly derived from the most valuable VASARI criteria (Jordan and Flanders 2019). These agreed elements would include terminology and numerical grading systems, and could also incorporate quantitative imaging biomarkers extracted by imaging post-processing software. A standardised approach also

has secondary benefits to teaching, research and quality improvement (Flanders and Lakhani 2012).

To determine how imaging studies are currently being reported for patients with glioma, this chapter will assess free-text reporting by neuroradiologists at a tertiary centre against a standardised set of reporting criteria derived from VASARI. The aim is to identify which information is most commonly missed and, upon a second read of the images, to also establish whether any potentially important features were missed in the free-text reports that would inform the ongoing care of the patient.

A wide variety of reporting style and content is anticipated, and by comparing free-text report content to a standardised reporting method based on VASARI, important missing information may be identified. Within the context of MRI biomarker translation for clinical use that I set out in Chapter 2, this study focuses on steps 1 and 2 of the translational framework, namely establishing the clinical use case and identifying potential imaging features of interest for translational development of quantitative imaging biomarkers for glioma reporting. I anticipate that it will inform the design of a useful reporting structure for standardised glioma reporting, which may combine structured report and quantification elements, with a view to facilitate both effective communication with the multidisciplinary team and patient management.

6.2 Methods

6.2.1 Case selection

Current glioma reporting in a tertiary neuroradiology centre (Lysholm Department of Neuroradiology, National Hospital for Neurology and Neurosurgery, Queen Square, London, UK) was assessed. Local clinical service improvement approval was given for this study. A hundred baseline reports were retrospectively identified for adult patients (>18 years old) with previously unknown gliomas that were either confirmed as glioma on histopathology or in the absence of tissue confirmation were judged to be glioma by multidisciplinary team meeting (MDT) consensus. These 100 cases were identified by review of neuro-oncology MDT meeting lists. Reports were all issued by a consultant neuroradiologist, including from those who provide the out-of-hours service. A consultant neuroradiologist was either the sole author of the report or was jointly reporting with a specialist neuroradiology trainee.

All reports were included based on the imaging protocol including pre- and post-gadolinium contrast enhanced T1-weighted, T2 weighted, FLAIR and DWI sequences. There were some cases where the patient has undergone two separate scans at baseline, initially without contrast and then contrast-enhanced, in very short succession. These were recorded as jointly reported by two consultants and the free text of each report was combined to incorporate all features mentioned. Additional information recorded was whether the patient has had a CT scan directly prior to the MRI scan, and any histopathology results if available.

6.2.2 VASARI scoring criteria

Free text reports were scored against a scoring proforma derived from the VASARI criteria. Table 6-1 sets out a brief summary of evidence to date that supports each VASARI feature. Features were grouped into the following categories: location; size; T2/FLAIR characteristics; contrast-enhanced T1 (T1+c) and diffusion characteristics; and all other features, as shown in Figure 6-1.

The scoring form was used to assess each of the 100 free-text reports. If a feature was mentioned, it was checked off on the form, with direct quotes from the reports recorded. If a reporter had documented that a feature was not present, in order to highlight an important negative finding (e.g., 'there is no calcification'), or if a feature was possible (e.g., 'there may be calcification'), this was also recorded.

Table 6-1. List of VASARI features and a brief description of their clinical utility.

VASARI Feature	*Evidence*
Tumour location (f1)	High agreement between raters (Gutman et al. 2013; Hyare et al. 2019)
Side of lesion centre (f2)	Highest agreement of the VASARI criteria, LGG subtypes can show differences in spatial distribution (Wijnenga et al. 2019)
Eloquent brain (f3)	May be associated with IDH mutation status (Hyare et al. 2019)
Enhancement quality (f4)	Can be helpful in distinguishing typical IDH^{wt} glioblastoma (rim enhancement) from IDH-mutant features (e.g. solid, speckled) (Berberich et al. 2018)
Proportion enhancing (f5)	Variably reported to be valuable for genotyping (Su et al. 2019; Zhou et al. 2017) and predictive of outcomes (Wangaryattawanich et al. 2015)
Proportion nCET (f6)	Potentially useful biomarker for IDH status in glioblastoma (Lasocki et al. 2017)
Proportion necrosis (f7)	Significant agreement between raters (Gutman et al. 2013)
Cysts (f8)	Useful for prediction of IDH mutation (Maynard et al. 2020).
Multifocal / multicentric (f9)	Multifocality associated with significantly worse prognosis for glioblastoma (Patil et al. 2012) and IDH1 mutation in LGG (Park et al. 2018).
T1/FLAIR ratio (f10)	Proportion nCET easier to record and more commonly reported. May predict mutation status (Hyare et al. 2019)
Thickness of enhancing margin (f11)	Difficult for human eye to accurately and consistently measure. Could be a useful genetic discriminating feature in diffuse midline gliomas (Chauhan et al. 2021).
Definition of enhancing margin (f12)	Not a widely investigated or useful sign in a recent systematic review (Lasocki et al. 2020)
Definition of non-enhancing margin (f13)	High agreement between raters, and reported as a predictor in an IDH typing study (Darlix et al. 2017).
Proportion oedema (f14)	Not possible to reliably distinguish oedema and non-enhancing infiltrative glioma components (Eidel et al. 2017)
Haemorrhage (f16)	Can be difficult to reproduce between raters and/or distinguish from mineralisation. Some evidence of association with 1p19q codeletion (Lasocki et al. 2020).
Diffusion characteristics (f17)	Can predict IDH mutation status (Xing et al. 2017) and differentiate HGG and LGG (Zhang et al. 2017).
Pial invasion (f18)	Possible prognostic differentiator in IDH wild type LGG (Park et al. 2020)
Ependymal extension (f19)	Associated with poorer outcomes for GBM (Mistry et al. 2017)
Cortical involvement (f20)	Low reader agreement. Possible prognostic indicator in LGG (Park et al. 2020)
Deep white matter invasion (f21)	Difficult to be sure of whether this is present on structural MRI but suspicion of major tract involvement could inform advanced imaging. May be predictive of IDH mutation status (Hyare et al. 2019).

nCET crosses midline (f22)	May be associated with IDH mutation status (Shen et al. 2020).
CET crosses midline (f23)	May be associated with IDH mutation status (Shen et al. 2020).
Satellites (f24)	Captured under (f9)
Calvarial remodelling (f25)	Indolent glioma subtypes are usually recognised well enough without relying on this sign, not considered useful in (Hyare et al. 2019).

6.2.3 Second read of images and free text reports

Following the initial read of the free-text reports and recording of contents using the proforma, a 'second read' of the images was performed. The same proforma was used to record the findings of the second read. Where there was any doubt over a feature, these cases and their second read structured reports were reviewed with an expert consultant neuroradiologist for consensus. Original report contents as recorded in the proforma were then compared to the second read reports and discrepancies identified. A score was given to each case to reflect the findings of any differences between original and second reports: 1. Any differences unlikely to affect interpretation; 2. Some important differences, however unlikely to change interpretation; 3. Some important differences which may impact on scan interpretation.

Information	Location	Size	T2/FLAIR characteristics	T1+c/diffusion characteristics	Other features
Subject ID	Laterality	No size	Proportion nCET/oedema	Proportion enhancing	Cysts
				Enhancement quality	Calcification
Scan date	Epicentre	Descriptive	Definition of non-enhancing margin	Proportion necrosis	Haemorrhage
				CET contacts/crosses midline	Pial invasion
Reported by	Eloquent brain	1x	nCET crosses/contacts midline	Definition of enhancing margin	Ependymal extension
Previous CT	Multifocal	2x	T1/FLAIR ratio	Diffusion description	Cortical involvement
					DWM invasion
Histology	Number of lobes	3x		Diffusion quantification	Calvarial remodelling

Figure 6-1. Glioma report assessment proforma based on VASARI criteria. nCET = non-contrast-enhancing tumour. T1+c = T1 + contrast. DWM = deep white matter.

6.3 Results

6.3.1 Subject demographics

A hundred patients with glioma were included who had undergone a baseline MRI which detected a glioma between the years 2017 and 2021. Their age range was 17-87, mean (SD) 54.4 (16.8). Sixty-six of these patients had a CT scan of their head in the days immediately before their MRI. On histopathology, 66% were found to have an HGG (WHO grade III or IV), 14% LGG, and 20% had no histopathology (of this group, 14 were suspected HGG and 6 suspected LGG).

6.3.2 Radiology reports

Reports were from the previous five years: 3 from 2017, 7 from 2018, 41 from 2019, 41 from 2020 and 8 from 2021. Most reports were by 21 individual consultant neuroradiologists, who reported 82 cases in total. Joint reports with two consultants involved were issued in 12 cases. Six reports were provided by the out-of-hours consultant service. The number of reports issued by each of these 23 sources is depicted in Figure 6-2, which shows that there is a wide variation of how many reports were issues by each, ranging from one to 12.

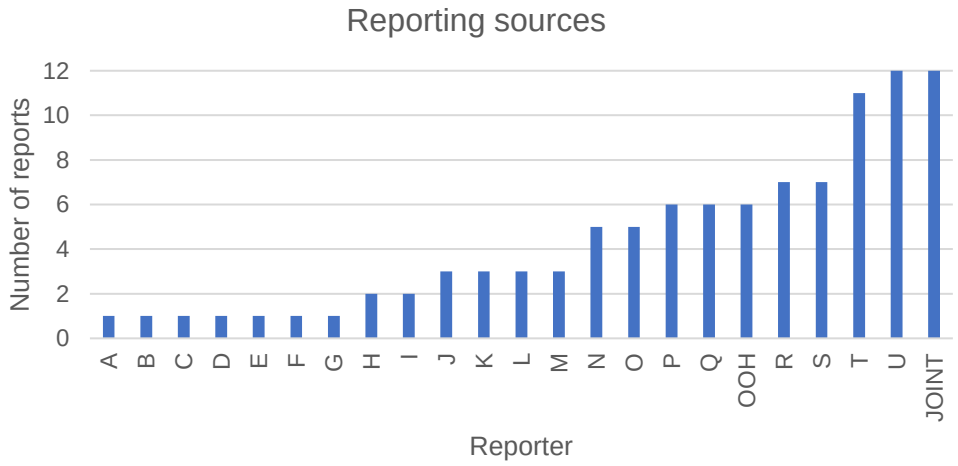


Figure 6-2. Bar chart showing the number of reports produced by each author. The letters signify individual consultant radiologists. OOH = out of hours service. JOINT = joint reports.

6.3.3 Features reported

The features that were included within the free-text reports by group of characteristics were recorded using the VASARI-based proforma (that is, location, size, T2/FLAIR, T1+c and diffusion, and other features). Graphs depicting the results of each category of characteristics as reported in read 1 and read 2 are displayed in Figure 6-3.

6.3.3.1 Location Characteristics

Almost all reports (99%) mentioned the laterality of the tumour and epicentre was recorded in 97%. Only 12% mentioned the involvement of eloquent brain regions. Twenty-one percent of reports stated that the tumour was either multifocal or multicentric with a further 5% stating that this was a possible feature and 43% stating that the tumour was not multifocal, as an important negative. The number of lobes affected was reported in 84% of cases.

6.3.3.2 Size characteristics

None of the reports included a volumetric measurement of the tumour. Only 20% provided measurement in three planes. Two percent included two plane measurement, 7% gave a measurement in one plane, and 33% used a descriptive word only (e.g. ‘large’ or ‘small’). The remaining 38% did not provide any measurement or description of tumour size, as shown in Figure 6-4.

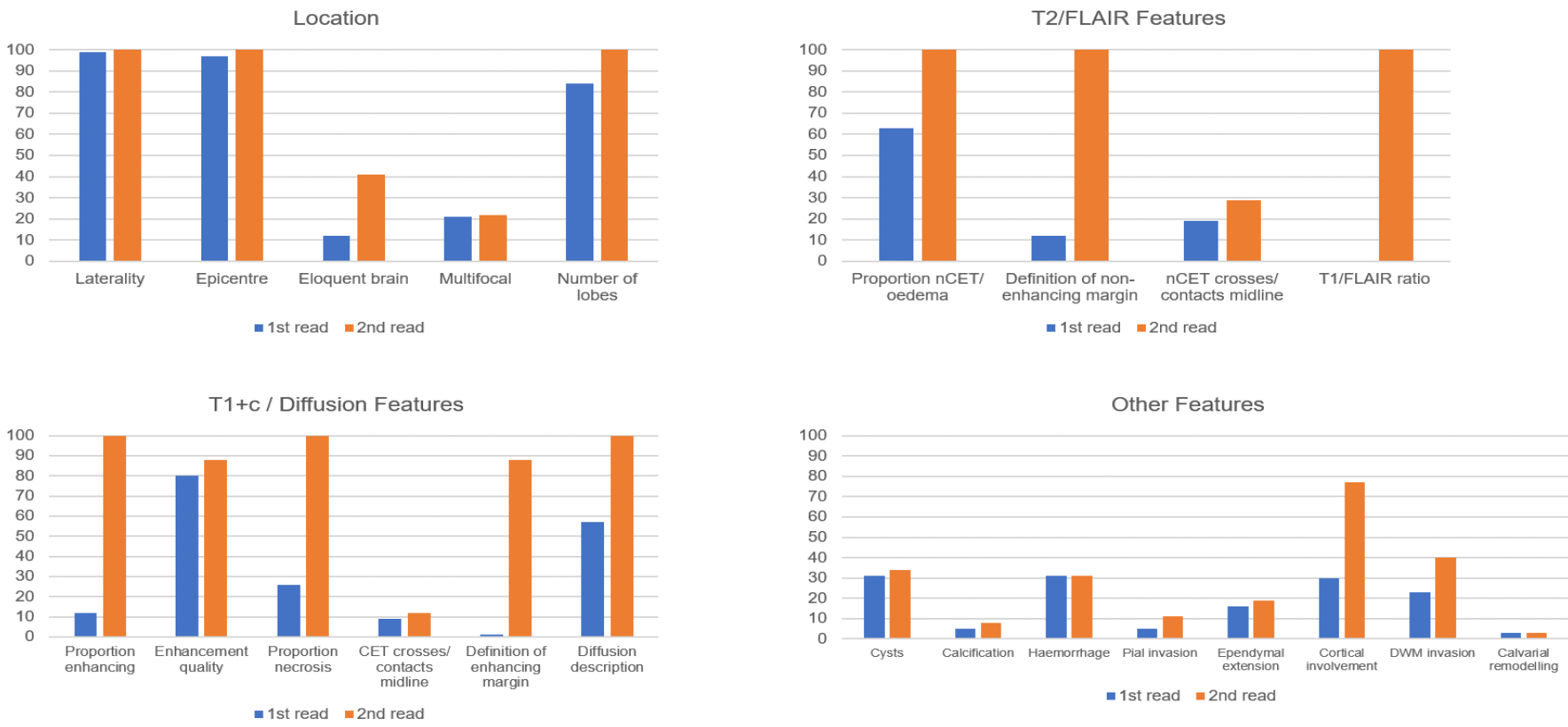


Figure 6-3. Results by set of features shown as bar graphs; free-text features '1st read' are shown in blue, and structured report features '2nd read' are shown in orange.

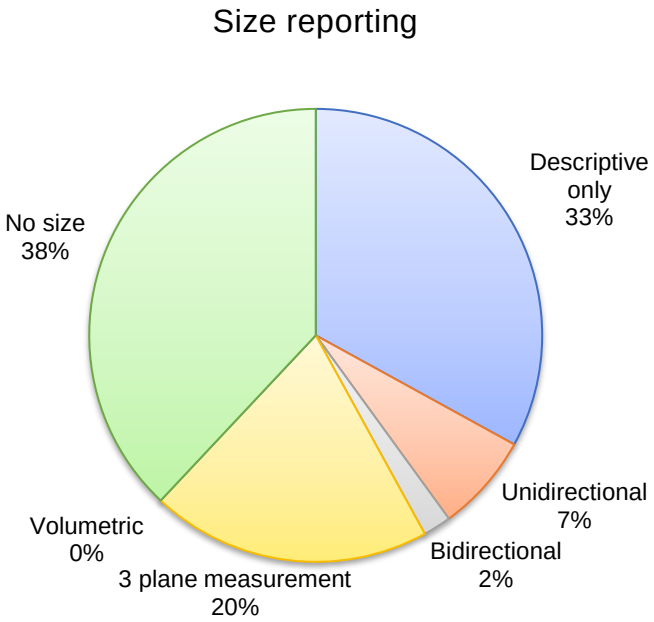


Figure 6-4. Pie chart showing how tumour size was reported.

6.3.3.3 T2/FLAIR characteristics

The proportion of nCET/oedema was not quantified in any of the reports. Thirty-five percent gave a qualitative assessment of the proportion however, using expressions like ‘there is marked surrounding oedema’. A further 28% mentioned that oedema was present without using a word to describe its proportion, and 2% mentioned that there was no oedema as an important negative (e.g., ‘there is no peritumoural T2/FLAIR hyperintensity’). The definition of the non-enhancing margin was only mentioned in 12% of cases, using descriptions such as ‘well defined’, ‘irregular’, and ‘heterogeneous’. Crossing or contact with the midline was mentioned by 19%, and a further 7% mentioned this as an important negative. The T2/FLAIR mismatch sign was mentioned more often as an important negative, in 13% of cases, and only as being present in 1% of cases. No reports mentioned the T1/FLAIR ratio.

6.3.3.4 T1+c and diffusion characteristics

Twelve percent of reports included a descriptive word denoting proportion of enhancement, using expressions like ‘there is moderate enhancement’. A further 14% mentioned that there was no enhancement. However,

substantially more reports included a description of the enhancement quality, at 80% of reports. Descriptions used included ‘patchy’, ‘peripheral irregular’, ‘homogeneous’ and ‘focal’. The definition of the enhancing margin is described by 1% of reports. Nine percent of reports state that the contrast enhancing tumour contacts the midline, and a further 2% include it as an important negative. The presence of necrosis with subjective description of its proportion was included in 26% of reports, using phrases like ‘predominantly necrotic’ and ‘small areas of necrosis’, and a further 1% including it as an important negative. Description of diffusion characteristics was more consistently included, by 57% of reports, using descriptions such as ‘reduced diffusivity’ and ‘central diffusion restriction’. In 12% this feature was included as an important negative (‘there is no restricted diffusion’). Diffusion was not quantified in any of the reports.

6.3.3.5 Other features

Cysts were mentioned as being present in 31% of reports and as possible in 1%. Calcification was reported as present in 5% of cases, being possible in another 5%, and absent in 3%. Haemorrhage was mentioned as being present in 31% reports, with a further 5% saying it was possible but uncertain, and 11% including it as an important negative. Pial invasion was reported to be present in 5% of reports, possible in 1% and absent in 7%. Ependymal extension was mentioned as present in 16% reports, and 3% included the feature as an important negative. Cortical involvement was mentioned by 30% of reports, as possible by 1% and as an important negative by 2%. Deep white matter invasion was reported in 23% of cases, as possible in 1% and as an important negative in 1%. Calvarial remodelling was a feature mentioned in 3% of reports, with a further 1% mentioning it as an important negative.

6.3.4 Second read of images

6.3.4.1 Location characteristics

All cases were described in terms of their laterality, epicentre, and number of lobes affected. Forty-one percent were recorded as involving eloquent brain regions as defined by the VASARI features (that is, involving one of speech motor, speech receptive, motor and vision anatomical regions). This was in

contrast to the 12% of cases where this was mentioned in the original reports. A similar number of cases to the original reports were described as multifocal, at 22%.

6.3.4.2 Size characteristics

All cases received a measurement of the tumour in three planes, as per the VASARI definition.

6.3.4.3 T2/FLAIR characteristics

All cases were described in terms of proportion of nCET/oedema and the definition of their non-enhancing margin. Twenty-nine percent of cases were reported to cross or contact the midline, in contrast to 19% of first reports.

6.3.4.4 T1+c/diffusion characteristics

All cases were described in terms of proportion enhancing, proportion necrosis, and their diffusion characteristics. Enhancement quality and definition of the enhancing margin was described in 88% of cases. Enhancing tumour crossing or contacting the midline was reported in 12% of cases, slightly more than the first reads which totalled 9%.

6.3.4.5 Other features

Cysts were reported in 34% of cases, slightly more than the 31% in first reports. Likewise for calcification, 8% of cases were reported as being positive for calcification in contrast to 5% of first reports, with a further 5% reported as having possible calcification. Haemorrhage was reported in 31% of cases and possible in 5%, which exactly matched first report frequencies. Pial invasion was reported in 11% of cases and possible in 6%, whereas it was only mentioned in 5% of first reports. Ependymal extension was reported in 19% and possible in 8%, which had been reported as present in 16% of first reads. Cortical involvement was found in 77% of cases, whereas this had only been mentioned in 30% of first reads. Forty percent of cases were reported to involve deep white matter tumour invasion, greater than the 23% mentioned in first reports.

6.3.5 Outcome grading

A pre-defined grading system was applied to each case by comparing the contents of first read and second read reports, an example of which can be seen in Table 6-2 and Figure 6-5. The large majority of cases, 82%, were assessed as grade 1, i.e. any differences would be unlikely to change overall interpretation. For all these cases there were additional features that were covered in the second read report, which contributed to a clearer picture of the imaging findings, however they did not cause a significant change in overall interpretation. These features included providing a three-plane measurement of tumour size.

For 12% of cases which were graded at level 2, the additional features included in second read reports were significant, however did not reach the threshold for affecting the overall interpretation of the scan. For example, these were cases where eloquent brain or deep white matter invasion was not described in the original report, which would be important for surgical planning, but the tumour was inoperable, or the patient was too frail to undergo surgery. The reporters that gave level 2 reports were: *H, I, M, OOH, R,* who with one report each at level 2; *U*, with two reports at level 2, and *JOINT* reports, with three reports at level 2.

Finally, 6% of cases were graded at level 3, where the important features that had been missed out of the first read reports may have had an impact on how the case would be interpreted. This included cases where involvement of eloquent brain or deep white matter was not described and it would be useful for surgical planning. Two cases underwent review of the images at MDT, where further functional MRI and diffusion tensor imaging tractography was recommended due to suspicion of involvement of eloquent areas. Two cases underwent emergency debulking before they could be reviewed at MDT. The final two cases were misreported as extra-axial. The reporters who gave level 3 reports were: *I* and *OOH* with one level 3 report each, and *N* and *U* with two level 3 reports each.

As numbers of report per reporter are small and there is a wide variety in the number of reports made by each, it is difficult to compare inter-rater performance. Comparing the three raters who wrote the most reports, *T, U* and *JOINT*, *T*'s reports were all at level 1, *U* had two reports at level 2 and two reports at level 3, and there were three *JOINT* reports at level 2.

Table 6-2. An example of first and second read comparison. This case was rated as level 3 – the extra features highlighted in red text were assessed to be significant to interpretation.

Feature	*Read 1*	*Read 2*
Laterality	Right	Right
Epicentre	Frontal	Frontal
Eloquent brain		Yes
Multifocal	No	No
Number of lobes	1	1
Size	No Size	3 planes
Proportion nCET/oedema	Moderate	50%
Definition of non-enhancing margin		Ill defined
nCET crosses/contacts midline		Yes
T1/FLAIR ratio		Mixed
Proportion enhancing		10%
Enhancing quality	Irregular	Marked/avid
Proportion necrosis	Large component	40%
CET contacts midline		No
Definition of enhancing margin		Well defined
Diffusion description	Free	Facilitated
Cysts	Yes	Yes
Calcification		No
Haemorrhage	Yes	Yes
Pial invasion		Yes
Ependymal extension		No
Cortical involvement		Yes
Deep white matter invasion		Yes
Calvarial remodelling		No

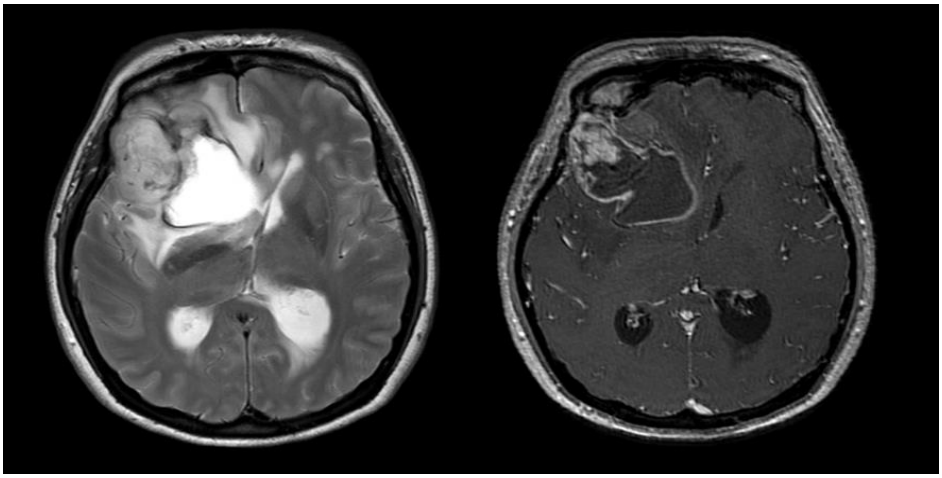


Figure 6-5. Axial T2 (left) and contrast enhanced T1 (right) images shown for the case reported in table 2.

6.4 Discussion

One hundred baseline clinical reports for glioma were assessed for their contents against the VASARI feature set. A second read of the imaging series provided a comparison. Specific features were found to be consistently under-reported in free text reports. These omissions fall into two broad categories.

The first category is expressing size or semi-quantitative properties. Tumour size was not reported at all in 38% of cases and was given a descriptive word in a further 33%. Additionally, there was a lack of reporting features of proportion, i.e. proportion of enhancement, necrosis, and oedema/infiltration. The second category relates to the documentation of detailed anatomical information. This includes missed descriptions of cortical involvement, possible deep white matter invasion, and the impact on eloquent brain regions.

Addressing these features in the baseline preoperative report may provide a more complete description that could facilitate multidisciplinary decision making and treatment planning. These decisions include whether the patient would benefit from any further imaging, for example if the baseline report raised the possibility of deep white matter tract or eloquent brain involvement, diffusion MRI and fibre tractography and functional MRI may be performed next. While these features are likely to be highlighted by the radiologist and neurosurgeon at an MDT review of the images, describing these features in the baseline report may mean that the patient undergoes advanced imaging more quickly, and that potentially all their imaging could be reviewed at the same MDT rather than having to be recalled prior to the surgical management plan being finalised.

By performing a third stage of analysis where first and second reads were compared and classified based on the degree of their discrepancies, there were cases identified where this type of additional information may have assisted in reaching management decisions. While no adverse outcomes were associated with missing information in baseline reports, since all cases were reviewed at MDT meetings and discussed by experts, conversely it was deemed that the omitted information may have facilitated or eased the pathway of patient management had it been present. This is difficult to define and is an

inherently a subjective judgement, however it can be supported in all cases by evidence from the patient's subsequent management records. For example, the patient underwent more advanced imaging following an MDT, when the anatomical features raising suspicion of involvement of eloquent brain areas had not been mentioned in the report. Another example is of a tumour being mistakenly reported as a likely meningioma, where more careful review of features such as pial invasion and cortical involvement would have helped to classify the tumour as intra-axial.

A larger minority of cases were classified as grade 2, where omissions were important but deemed not to meet a threshold of impacting patient management. This category included cases where the omissions were equivalent to grade 3 cases in terms of their potential significance, but where the patient was significantly frail, surgical management was not considered viable and instead the patient received best supportive care. In other grade 2 cases the tumour itself was deemed inoperable due to its size or location, and this mitigated the importance of detailed feature description. There were few cases where there was a significant straightforward human error, which ultimately could be corrected and would be unlikely to affect patient management, for example where the radiologist described the tumour as being on the left when it was actually on the right. These human error events may be reduced by use of structured reporting. Since the sample used contained reports by so many different radiologists and since the split between them was uneven it is difficult to make conclusions about inter-rater performance. The six cases with level 3 errors were made by four reporters who had reported between 2 and 12 cases. Of the twelve joint reports, three contained level 2 errors.

6.4.1 Structured reporting

Structured reporting offers several other potential benefits in the context of clinical glioma reporting. In free-text reports, the absence of certain features is inconsistently mentioned by reporters in order to highlight these features as important negatives. It is clear from the results that important negatives are very inconsistently reported. The presence of every feature on a structured

report means that by default no important negatives would go unmentioned. Structured reporting would also encourage consistency in feature reporting by prompting the reporter for clear concise descriptions. Report design could specify that a reporter provides a two-plane measurement of the tumour and even provide brief instructions for which planes and maximum diameters to select. However, the correct balance would need to be achieved. It would be important to provide the reporter with the opportunity to combine structured and free-text descriptions, since not all features of an imaging examination can be captured by the VASARI features. Several examples of this important additional information were encountered within this dataset, for example the description of a previous stroke, the encroachment of tumour on a major blood vessel, and crucially several cases where important descriptions were given of midline shift and/or other worrying signs of raised intracranial pressure which necessitated urgent neurosurgical review.

Certain VASARI features are associated with a paucity of evidence or are difficult to fulfil, for example providing a measurement of thickness of tumour margin. Calvarial remodelling is of limited importance as an indicator of an indolent LGG. Therefore the structured report itself may benefit from refining, taking into account the features that are both the most relevant and also the most under-reported and focusing on those, in order to make the exercise as beneficial as possible, while increasing reporter engagement and reducing reporting time.

Structured reports have been shown to increase clinical referrer satisfaction due to content and report clarity (Schwartz et al. 2011) and reduce feature omissions (Lin, Powell, and Kagetsu 2014). This may be particularly relevant for glioma reporting given the wide range of possible features and appearance heterogeneity, a radiologist may focus on the same few features for every report or miss important additional features due to ‘satisfaction of search’ (Ganeshan et al. 2018). A careful balance must be reached between adopting structured reporting for its benefits and still ensuring that the radiologist is able to fully express their impressions without introducing perceived or actual limitations (Weiss and Langlotz 2008).

The baseline reporting template designed by the brain tumour reporting and data systems (BT-RADS) focuses on a sub-selection of VASARI features: tumour location, FLAIR abnormality, enhancement, and diffusion properties (www.btrads.com, BTRADS 2021). It also signposts additional features to check including whether there is any evidence of infarction, hydrocephalus or significant haemorrhage. Its follow-up report includes a progression score based on the Response Assessment in Neuro-oncology (RANO) criteria (Wen et al. 2010). BT-RADS reports have been shown to be more concise and include less ambiguity than free-text reports (Zhang et al. 2019), however their accuracy and completeness have not been compared to reports containing a more extensive representation of the VASARI features.

6.4.2 Glioma imaging biomarker quantification

Additionally, the potential role of quantification or semi-quantification as part of a structured clinical reporting system has not been widely addressed. In a survey of 220 European radiology centres, it was clear that very few centres used quantification methods to assess parameters like tumour size or ADC values (Thust et al. 2018). This may be due to a combination of factors including a lack of available software, limited opportunities for user training, and reporting time pressures. While there is RANO guidance recommending two-plane measurement of tumour size at follow up, most radiologists provide a visual descriptive estimate, a finding that has been reproduced in this study. Two-plane measurement has been shown to correlate well with tumour volume, particularly in HGGs (Galanis et al. 2006). Another study showed that two-plane measurement had moderate intra-rater agreement but that a tumour had to have grown by 10ml for a radiologist to detect that it had progressed (Berntsen et al. 2020).

Many key glioma features lend themselves to quantification, but as demonstrated in this study these are some of the features that are consistently under-reported. Tumour irregularity and infiltration combined with the overall heterogeneity of glioma features means that manual measurement of quantitative features can be highly variable both within and between raters (Bø et al. 2017; Vos et al. 2003).

Automated tumour segmentation methods could therefore have an important clinical role, potentially providing whole tumour volumetry as well as tissue composition information by segmentation of necrotic, enhancing and oedema/infiltration components. Algorithms continue to be technically validated against each other in the research setting, commonly through initiatives like the annual Brain Tumour Image Segmentation (BRATS) benchmark challenge (Menze et al. 2015). Deep learning algorithms have shown technical promise with standardised research-quality data (Pereira et al. 2016) and less frequently with smaller cohorts of clinical grade data at baseline and for longitudinal analysis (Meier et al. 2016; Porz et al. 2014). Efforts towards broader feature extraction for HGG neurosurgical planning, including distance or overlap of tumour with particular brain structures and other VASARI features, have shown promise in a recent large multi-centre study (Kommers et al. 2021). Further clinical validation is needed to demonstrate that automated deep learning-based segmentation and feature extraction can perform reliably across glioma grades despite clinical challenges such as robustness to multiple scanners, acquisitions, missing sequences, and computational constraints.

6.4.3 Towards a quantitative report for gliomas

This study has been focused on the early stages of imaging biomarker translation as per the framework I described in Chapter 2. It will contribute to setting the clinical picture for further work towards introducing automated reporting for gliomas and performing clinical validation in a similar way as I have for other disease areas. A structured quantitative report for glioma should focus on the features that are difficult to describe, under-reported, and known to be of diagnostic and management benefit, which have been highlighted by this study. It would require robust technical validation to select a segmentation technique that is shown to perform optimally with heterogenous clinical data, application of clinically useful brain atlases, and demonstration of accurate registration and labelling in the presence of anatomical distortion.

Standardisation and consistent reporting of key glioma imaging biomarkers in the clinical setting has the potential to facilitate streamlined patient

management in terms of neurosurgical and/or radiotherapy planning; improve communication between clinicians; contribute towards training of radiologists and neurosurgeons; promote adoption of precision medicine and provide a rich source of clinical data for radiogenomic analysis (Smits 2021).

7. Conclusions

7.1 Thesis overview

7.1.1 Summary of work

Having identified that quantitative imaging biomarkers (QIBs) face many barriers to clinical adoption and often lack adequate clinical validation, I began by setting a practical framework for the development of quantitative reporting tools to promote their effective translation to the clinical neuroradiology setting. The framework I set out places clinical validation at the centre of the translational process, highlighting the importance of both pre-use validation and in-use evaluation. Reframing the technical development of quantification methods with a clinical application in mind from the outset is likely to foster the development of validated tools that will serve an outstanding clinical need, encourage clinician engagement, and prioritise workflow integration.

I applied elements and principles from the translational framework to the development of quantification tools for several disease areas. For hippocampal sclerosis (HS) as a cause of focal epilepsy, technically validated methods for hippocampal volume and signal intensity measurement were combined and used to process normative reference data, from which a quantitative report was constructed. I designed and performed a clinical validation study involving multiple clinical end-users with different experience levels. A study that combines visual assessment with quantitative information has not previously been performed for HS in the literature. This study showed that when users integrated the information from the quantitative report into their assessments, they were more accurate in detecting HS, especially cases of bilateral HS.

Improved detection of HS has positive implications for clinical decision making, in particular bilateral HS non-curative surgical attempts may be reduced. Report users were also more confident in their correct assessments than when it was not available. Demonstrating the clinical validity and potential added value of the HS quantitative report has paved the way for its clinical workflow integration into the local neuroradiology department, and in-use evaluation is under way. Positive engagement of radiologists, radiographers and clinical scientists across the department is facilitating the integration process.

A similar clinical validation approach was applied to a volumetric quantification tool for use in suspected dementia. The clinical validation study I conducted that compared reporters' accuracy and confidence with and without the quantitative report available showed that reports conferred increased sensitivity to detection of abnormal volume loss and subjects with Alzheimer's Disease. The most experienced raters also displayed increased accuracy in their assessments and increased agreement with the gold standard. Inter-rater agreement also improved with the report from good to excellent agreement. The results for this clinical validation study reflect the complexity of dementia imaging assessment: regional brain volume measurements convey complex patterns associated with different pathologies that can overlap with normal ageing appearances in the early stages. This reinforces the principle that quantitative reports should be used as complementary additional information for the radiologist to assimilate into their overall assessment. While there are many commercial tools available for use in dementia imaging assessment, none have published a clinical validation study of similar scope, the largest published involving just two raters. Plans that I set out for a larger validation study engaging even more raters would build on the positive foundations of this work.

For quantification of brain volumes and lesions in multiple sclerosis, I tackled a different translational challenge, that of applying a processing pipeline that usually requires non-protocol T1-weighted images to widely clinically available FLAIR images. I showed that using FLAIR as the single pipeline input produced comparable results to the conventional multi-sequence scenario in a multi-centre, multi-vendor clinical dataset of patients with clinically isolated syndrome. The study allowed me to explore additional translational challenges, including the need to promote acquisition standardisation and account for field strength differences. I built on the positive results of the FLAIR-only pipeline by constructing a normative reference dataset that covers a wide age range, which will be used as the foundations of a quantitative MS reporting tool. It will be more directly applicable to the routine clinical setting than available commercial tools, which have not dealt with the significant translational barrier of requiring multi-sequence inputs.

For gliomas, the focus was on gaining an understanding of current clinical reporting practices and using this to establish the possible clinical need for development of a quantitative reporting tool. This study established that the content of free-text reports is highly variable, and that they often lack tumour measurements and other potentially important anatomical and location details that are considered when planning treatment. Some reports reviewed had omissions that had the potential to impact on management decisions if the reports were used in isolation of multidisciplinary review meetings. The important missing features identified have the potential to be automatically quantified and this work paves the way towards constructing a clinically meaningful quantitative report for glioma imaging.

7.1.2 Further applications

There are numerous interesting extensions leading from the work I have presented in this thesis within the same disease areas. For example, the HS report could be modified to quantify FLAIR signal properties instead of T2, which would make it more generalisable to centres which do not routinely perform T2 relaxometry. A longitudinal report of hippocampal atrophy and signal change over time may be useful in cases where surgical resection is not immediately indicated, for example where changes are very subtle at symptom onset and clinicians would prefer to repeat imaging after a time interval to assess for developing HS imaging appearances. There are also possibilities to explore detection methods for other causes of focal epilepsies, for example for focal cortical dysplasia (FCD), where there are several different techniques available which assess for abnormalities of the cortical grey matter (Wang, Ahmed, and Mandal 2020).

For dementia, a natural progression from the cross-sectional reporting tool would be to develop a report for longitudinal comparisons. This would require extension of the normative reference dataset to include longitudinal follow-ups of healthy controls so that rates of atrophy could be meaningfully compared. A reference AD population could also be included, so that an individual's atrophy rate can be plotted against both populations. Since the current reporting tool processes the T1-weighted series only to quantify brain volumes, a useful

extension would be to extend the processing pipeline to include other sequences for assessment of the white matter. Assisted detection of white matter hyperintensities, as well as microbleeds and lacunes, would provide a more complete picture of the impact of vascular disease or assist with the assessment of potential vascular dementia.

The MS FLAIR-only processing pipeline could usefully be adapted to provide a longitudinal assessment for change in lesion load and provide a measurement of an individual's brain atrophy rate. This would similarly require additional longitudinal reference data for healthy control and MS populations to be processed. As discussed in chapter 5, brain age modelling could form a novel extension to a quantitative MS reporting tool.

The work on current glioma reporting informs quantitative feature extraction for a glioma reporting tool. Longitudinal assessment is also important in glioma imaging, particularly for low grade gliomas to assess for high-grade transformation. Inclusion of QIBs such as diffusion values or perfusion measurements in a quantitative report for glioma would require careful validation and attention paid to result interpretability for the individual subject.

7.2 Outlook

Clinical application of radiological quantification solutions is still in its infancy. Commercial offerings of automated quantitative analysis tools are rapidly increasing, employing both artificial intelligence (AI) and non-AI based solutions (Rezazade Mehrizi, van Ooijen, and Homan 2021). Validation of these tools has not kept up with their development and in most cases their efficacy and value are still unknown. The FDA has produced an action plan in response to this growth period, providing guidance for development of AI-based software as a medical device (fda.gov/media/145022/download, FDA 2021). They set out plans to update existing regulatory frameworks and state that development will be encouraged to take a patient-centred approach and promote real-world evaluation studies. Similarly, a group of academic and commercial stakeholders have recently published the ECLAIR guidelines (Omoumi et al. 2021), which guides intended users and institutions on how to evaluate available AI solutions when considering their adoption. They place emphasis on the key elements of validation, usability, and integration that I have been exploring in this thesis.

Despite the availability and increasing uptake of automated quantification tools, there is no publicly accessible data available on the in-use evaluation of commercial quantitative reporting tools once they have been adopted into a clinical environment. European Medical Devices Regulation have recently declared that companies must start to report post-market clinical follow-up which includes data on the tool's clinical performance and safety data to achieve certification or revalidation, so by necessity it is likely that this will become more transparent in the future (ec.europa.eu/health/md_sector/overview_en, EU 2021).

Development of quantitative reporting should engage multiple stakeholders to work together to embed these new opportunities into clinical radiology. This multi-stakeholder approach, where partnerships are formed between academia, healthcare institutions and commercial organisations, should in theory create an environment where products are developed with attention paid to technical robustness and clinical applicability (Recht et al. 2020). Each

stakeholder may contribute different areas of expertise and access, which complement each other to assist clinical translation, as illustrated in Figure 7-1.

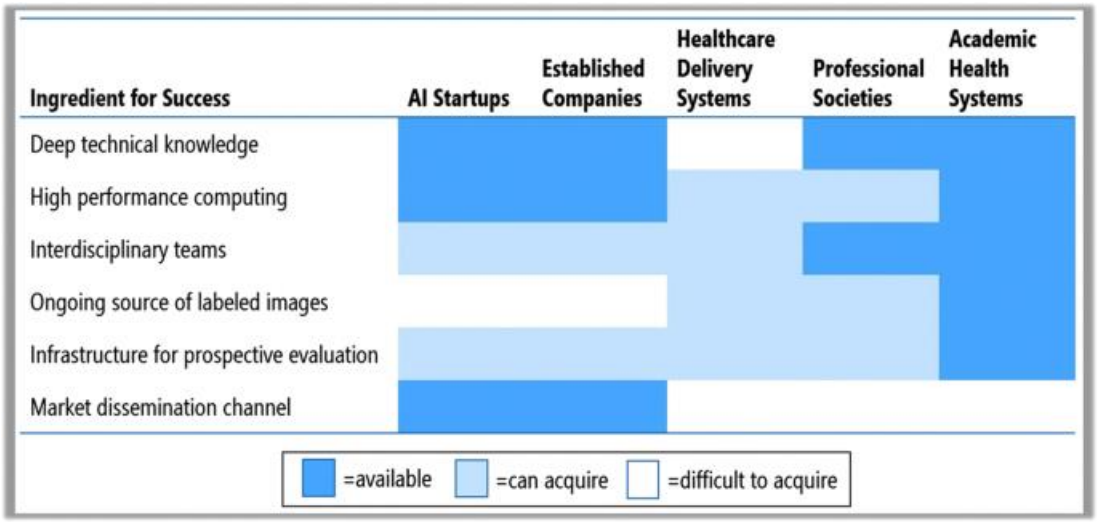


Figure 7-1. Stakeholders in the development of new quantification tools for clinical application can offer different strengths. Figure from Recht et al. 2020*.*

Engagement of radiologists as key stakeholders is crucial for successful clinical translation of quantitative reporting tools. There is a spectrum of attitudes towards integration of these tools into their reporting workflows, ranging from scepticism and mistrust to enthusiasm and high expectations that may be difficult to fulfil (Strohm et al. 2020).

For radiologists to overcome their scepticism or mistrust of these applications, or indeed to dispel the perception that the tool may have an unrealistic positive impact on their work, it has been proposed that some training in the principles of image processing and analysis methods should be incorporated into radiology training programmes and examinations (Recht et al. 2020). A recent international survey showed that while approximately half of the 1041 respondents had an open and proactive attitude, radiologists who only had a basic understanding of these concepts were significantly more fearful of job replacement and less open to adopting them (Huisman et al. 2021).

Educational resources are being established by, for example, the American College of Radiology's Data Science Institute (acrdsi.org, ACR 2021), and the European Society of Radiology (myesr.org/ai, ESR 2021a). Technical education should be accompanied by scientific evidence that supports implementation of the new tool in the form of clinical validation studies. As

discussed, this evidence is currently sparse, and presents a significant implementation barrier for radiologists and referring clinicians who value evidence-based healthcare (Strohm et al. 2020).

A recent review of CE marked AI image analysis applications across radiology evaluated all relevant published literature on their validation using a hierarchical model of efficacy, presented in Table 7-1 (van Leeuwen et al. 2021), which aligns closely with elements of the QNI framework.

Table 7-1. A hierarchical model for the level of evidence of efficacy of an imaging tool. From van Leeuwen et al. 2021.

Level	Explanation	Typical measures
Level 1t	Technical efficacy Article demonstrates the technical feasibility of the software	Reproducibility, inter-software agreement, error rate
Level 1c	Potential clinical efficacy Article demonstrates the feasibility of the software to be clinically applied	Correlation to alternative methods, potential predictive value, biomarker studies
Level 2	Diagnostic accuracy efficacy Article demonstrates the stand-alone performance of the software	Standalone sensitivity, specificity, area under the ROC curve, or Dice score
Level 3	Diagnostic thinking efficacy Article demonstrates the added value to the diagnosis	Radiologist performance with/without AI, change in radiological judgement
Level 4	Therapeutic efficacy Article demonstrates the impact of the software on the patient management decisions	Effect on treatment or follow-up examinations
Level 5	Patient outcome efficacy Article demonstrates the impact of the software on patient outcomes	Effect on quality of life, morbidity, or survival
Level 6	Societal efficacy Article demonstrates the impact of the software on society by performing an economic analysis	Effect on costs and quality-adjusted life years, incremental costs per quality-adjusted life year

Of the 100 commercial tools they identified, the authors found that 64 had no published evidence at any level. Twenty-seven percent of the studies published relating to the other 36 products included some assessment at level 3 efficacy or higher. For neuroradiology applications, this level of efficacy was only reached by two studies, which have been discussed elsewhere in this thesis (Brewer 2009; Ross et al. 2015). Only a minority of identified studies used multi-centre and multi-vendor data.

Radiologists occupy the forefront of the digitised work environment before all other healthcare disciplines, and many identify themselves as the logical pioneers for use of automated support tools (Strohm et al. 2020). They should be actively engaged in the development of these tools, as domain experts who can identify areas of clinical need, as well as conceptualise and design solutions with a higher likelihood of successful clinical translation (Scheek, Rezazade Mehrizi, and Ranschaert 2021). Their engagement is also key for

managing integration of reporting workflows, including coordination with radiographers and PACS managers. In-use evaluation by early adopters allows for common issues to be identified and for testing of the tool on complex or unusual cases, thereby more clearly defining the clinical use boundaries of the software solution.

Demonstrating health economic benefit of a quantitative tool is a fundamental aspect of assessing its clinical impact, however this is difficult to demonstrate. A recent paper estimated the average cost per patient for volumetric quantification in neurocognitive diseases to be approximately eighty-two US dollars (Raji, Ly, and Benzinger 2019) when using regulatory-approved commercial solutions. It is difficult to establish how much costs may vary depending on operational and implementation factors as well as nuances of a particular healthcare system and reimbursement strategies. Actual final cost is much more difficult to calculate in terms of impact on an individual patient's healthcare expenditure and potential added value. Health technology assessment (HTA) is a multidisciplinary process that uses explicit methodology to determine the value of a given health technology, taking into account existing alternatives, clinical effectiveness, safety, and economic implications, as well as ethical and legal issues, organisational aspects, and wider impact on patients and the population (Drummond et al. 2008; O'Rourke, Oortwijn, and Schuller 2020). In the United Kingdom, the National Institute for Health and Clinical Excellence (NICE) uses HTAs to evaluate health technologies for use in the National Health Service (NHS), within its evidence standards framework for digital health technologies (NICE 2018).

As current development of quantitative reporting tools tends to be in silos, with each solution being designed towards a specific task or disease area, costs and infrastructure demands on a radiology service to be able to implement several of them at once may become inhibitive (van Leeuwen et al. 2021). The development of generalisable algorithms and suites of applications that may become available directly from scanner manufacturers or PACS companies may be valuable as more tools become clinically validated.

More attention needs to be directed towards availability and quality of reference data. As discussed in previous chapters, there are several international research-led initiatives towards image acquisition standardisation and data sharing. As more clinical images start to enter the quantification pipeline, within due boundaries of ethics and data sharing regulations, their quantitative results could form a rich source of reference data for directly relevant clinical quantification tools as well as research into the rapidly expanding field of radiomics (Hosny et al. 2018). Ultimately radiomics aims to incorporate QIB signatures with other disease biomarker and patient demographic characteristics to deliver personalised diagnostic and prognostic assessment. Solutions for image optimisation and data harmonisation will play an important role in utilising clinical data for big data analysis and construction of multi-site reference populations (Fortin et al. 2018).

7.2.1 Summary

In this thesis I have established that quantification tools for radiology reporting should be viewed as adjuncts to the radiologist's expert visual assessment. They should be developed with contributions from multiple stakeholders all working towards delivering a technically robust, clinically validated solution that can be integrated into clinical workflows. Translational barriers exist that require companies, research bodies and clinicians to work together in partnership to overcome. These include application generalisability, image acquisition standardisation, and sourcing of relevant reference data. The potential value of these tools to radiologists, and subsequently to their patients, needs to be underpinned by accessible transparent evaluation of their scientific validity, with increased priority given to the demonstration of clinical assessment accuracy and proactive evaluation of their performance once they are adopted into the clinical environment.

8. References

ACR. 2021. “Data Science Institute DSI | American College of Radiology.” Retrieved August 23, 2021 (https://www.acrdsi.org/).

ADMdx. 2021. “ADMdx: Turning Images into Insights.” Retrieved August 23, 2021 (https://admdx.com/).

ADNI. 2021. “ADNI | ADNI 3.” Retrieved August 23, 2021 (http://adni.loni.usc.edu/adni-3/).

Ahmadi-Abhari, Sara, Maria Guzman-Castillo, Piotr Bandosz, Martin J. Shipley, Graciela Muniz-Terrera, Archana Singh-Manoux, Mika Kivimäki, Andrew Steptoe, Simon Capewell, Martin O’flaherty, and Eric J. Brunner. 2017. “Temporal Trend in Dementia Incidence since 2002 and Projections for Prevalence in England and Wales to 2040: Modelling Study.” *BMJ (Online)* 358.

Arbabshirani, Mohammad R., Sergey Plis, Jing Sui, and Vince D. Calhoun. 2017. “Single Subject Prediction of Brain Disorders in Neuroimaging: Promises and Pitfalls.” *NeuroImage* 145(Pt B):137–65.

Asao, Chiaki, Toshinori Hirai, Shunji Yoshimatsu, Tetsuya Matsukawa, Masanori Imuta, Katsuro Sagara, and Yasuyuki Yamashita. 2008. “Human Cerebral Cortices: Signal Variation on Diffusion-Weighted MR Imaging.” *Neuroradiology* 50(3):205–11.

Azab, M., M. Carone, S. H. Ying, and D. M. Yousem. 2015. “Mesial Temporal Sclerosis: Accuracy of NeuroQuant versus Neuroradiologist.” *American Journal of Neuroradiology* 36(8):1400–1406.

Barentsz, Jelle O., Jeffrey C. Weinreb, Sadhna Verma, Harriet C. Thoeny, Clare M. Tempany, Faina Shtern, Anwar R. Padhani, Daniel Margolis, Katarzyna J. Macura, Masoom A. Haider, Francois Cornud, and Peter L. Choyke. 2016. “Synopsis of the PI-RADS v2 Guidelines for Multiparametric Prostate Magnetic Resonance Imaging and Recommendations for Use.” *European Urology* 69(1):41–49.

Battaglini, Marco, Giordano Gentile, Ludovico Luchetti, Antonio Giorgio, Hugo Vrenken, Frederik Barkhof, Keith S. Cover, Rohit Bakshi, Renxin Chu, Maria Pia Sormani, Christian Enzinger, Stefan Ropele, Olga Ciccarelli, Claudia Wheeler-Kingshott, Marios Yiannakas, Massimo Filippi, Maria Assunta Rocca, Paolo Preziosa, Antonio Gallo, Alvino Bisecco, Jacqueline Palace, Yazhuo Kong, Dana Horakova, Manuela Vaneckova, Claudio Gasperini, Serena Ruggeri, and Nicola De Stefano. 2019. “Lifespan Normative Data on Rates of Brain Volume Changes.” *Neurobiology of Aging* 81:30–37.

Berberich, Anne, Thomas Hielscher, Philipp Kickingereder, Frank Winkler, Katharina Drüschler, Lars Riedemann, Marlene Arzt, Tobias Kessler, Michael Platten, Andreas von Deimling, Wolfgang Wick, Felix Sahm, Martin Bendszus, and Antje Wick. 2018. “Nonmeasurable Speckled Contrast-Enhancing Lesions Appearing During Course of Disease Are Associated With IDH Mutation in High-Grade Astrocytoma Patients.” *International Journal of Radiation Oncology, Biology, Physics* 102(5):1472–80.

Bernasconi, Andrea, Neda Bernasconi, Zografos Caramanos, David C. Reutens, Frederick Andermann, François Dubeau, Donatella Tampieri, Bruce G. Pike, and Douglas L. Arnold. 2000. “T2 Relaxometry Can Lateralize Mesial Temporal Lobe Epilepsy in Patients with Normal MRI.” *NeuroImage* 12(6):739–46.

Bernasconi, N., D. Kinay, F. Andermann, S. Antel, and A. Bernasconi. 2005. “Analysis of

Shape and Positioning of the Hippocampal Formation: An MRI Study in Patients with Partial Epilepsy and Healthy Controls." *Brain* 128(10):2442–52.

Berntsen, Erik Magnus, Anne Line Stensjøen, Maren Staurset Langlo, Solveig Quam Simonsen, Pål Christensen, Viggo Andreas Moholdt, and Ole Solheim. 2020. "Volumetric Segmentation of Glioblastoma Progression Compared to Bidimensional Products and Clinical Radiological Reports." *Acta Neurochirurgica* 162(2):379–87.

Biberacher, Viola, Paul Schmidt, Anisha Keshavan, Christine C. Boucard, Ruthger Righart, Philipp Sämann, Christine Preibisch, Daniel Fröbel, Lilian Aly, Bernhard Hemmer, Claus Zimmer, Roland G. Henry, and Mark Mühlau. 2016. "Intra- and Interscanner Variability of Magnetic Resonance Imaging Based Volumetry in Multiple Sclerosis." *NeuroImage* 142:188–97.

Bink, Andrea, Jan Benner, Julia Reinhardt, Anthony de Vere-Tyndall, Bram Stieltjes, Nicolin Hainc, and Christoph Stippich. 2018. "Structured Reporting in Neuroradiology: Intracranial Tumors." *Frontiers in Neurology* 9(FEB).

Bishop, Courtney A., Rexford D. Newbould, Jean SZ Lee, Lesley Honeyfield, Rebecca Quest, Alessandro Colasanti, Rehiana Ali, Miriam Mattoscio, Antonio Cortese, Richard Nicholas, Paul M. Matthews, Paolo A. Muraro, and Adam D. Waldman. 2017. "Analysis of Ageing-Associated Grey Matter Volume in Patients with Multiple Sclerosis Shows Excess Atrophy in Subcortical Regions." *NeuroImage: Clinical* 13:9–15.

Bø, Hans Kristian, Ole Solheim, Asgeir Store Jakola, Kjell-Arne Kvistad, Ingerid Reinertsen, and Erik Magnus Berntsen. 2017. "Intra-Rater Variability in Low-Grade Glioma Segmentation." *Journal of Neuro-Oncology* 131(2):393–402.

Bocchetta, Martina, M. Jorge Cardoso, David M. Cash, Sebastien Ourselin, Jason D. Warren, and Jonathan D. Rohrer. 2016. "Patterns of Regional Cerebellar Atrophy in Genetic Frontotemporal Dementia." *NeuroImage: Clinical* 11:287–90.

de Boer, Renske, Henri A. Vrooman, Fedde van der Lijn, Meike W. Vernooij, M. Arfan Ikram, Aad van der Lugt, Monique M. B. Breteler, and Wiro J. Niessen. 2009. "White Matter Lesion Extension to Automatic Brain Tissue Segmentation on MRI." *NeuroImage* 45(4):1151–61.

Boutet, Claire, Marie Chupin, Olivier Colliot, Marie Sarazin, Gurkan Mutlu, Aurélie Drier, Audrey Pellot, Didier Dormont, Stéphane Lehéricy, and Alzheimer's Disease Neuroimaging Initiative. 2012. "Is Radiological Evaluation as Good as Computer-Based Volumetry to Assess Hippocampal Atrophy in Alzheimer's Disease?" *Neuroradiology* 54(12):1321–30.

Braak, H. and E. Braak. 1995. "Staging of Alzheimer's Disease-Related Neurofibrillary Changes." *Neurobiology of Aging* 16(3):271–78; discussion 278-84.

Brainminer. 2021. "Brainminer - Machine Learning for Neurology." Retrieved August 23, 2021 (https://www.brainminer.co.uk/).

brainreader. 2021. "Brainreader - Neuroreader® Brain Scan and MRI Software." Retrieved August 23, 2021 (https://brainreader.net/).

Brewer, J. B. 2009. "Fully-Automated Volumetric MRI with Normative Ranges: Translation to Clinical Practice." *Behavioural Neurology* 21(1):21–28.

Brewer, J. B., S. Magda, C. Airriess, and M. E. Smith. 2009. "Fully-Automated Quantification of Regional Brain Volumes for Improved Detection of Focal Atrophy in Alzheimer

Disease." *American Journal of Neuroradiology* 30(3):578–80.

Briellmann, Regula S., Renate M. Kalnins, Samuel F. Berkovic, and Graeme D. Jackson. 2002. "Hippocampal Pathology in Refractory Temporal Lobe Epilepsy: T2-Weighted Signal Change Reflects Dentate Gliosis." *Neurology* 58(2):265–71.

Brinkmann, Benjamin H., Hari Guragain, Daniel Kenney-Jung, Jay Mandrekar, Robert E. Watson, Kirk M. Welker, Jeffrey W. Britton, and Robert J. Witte. 2019. "Segmentation Errors and Intertest Reliability in Automated and Manually Traced Hippocampal Volumes." *Annals of Clinical and Translational Neurology* 6(9):1807–14.

Brune, Synne, Einar A. Høgestøl, Vanja Cengija, Pål Berg-Hansen, Piotr Sowa, Gro O. Nygaard, Hanne F. Harbo, and Mona K. Beyer. 2020. "LesionQuant for Assessment of MRI in Multiple Sclerosis—A Promising Supplement to the Visual Scan Inspection." *Frontiers in Neurology* 0:1700.

BTRADS. 2021. "Home - Brain Tumor Reporting and Data System (BT-RADS)." Retrieved August 23, 2021 (http://btrads.com/).

Buchlak, Quinlan D., Nazanin Esmaili, Jean Christophe Leveque, Christine Bennett, Farrokh Farrokhi, and Massimo Piccardi. 2021. "Machine Learning Applications to Neuroimaging for Glioma Detection and Classification: An Artificial Intelligence Augmented Systematic Review." *Journal of Clinical Neuroscience* 89:177–98.

Carass, Aaron, Snehashis Roy, Amod Jog, Jennifer L. Cuzzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Ferran Prados, Carole H. Sudre, Manuel Jorge Cardoso, Niamh Cawley, Olga Ciccarelli, Claudia A. M. Wheeler-Kingshott, Sébastien Ourselin, Laurence Catanese, Hrishikesh Deshpande, Pierre Maurel, Olivier Commowick, Christian Barillot, Xavier Tomas-Fernandez, Simon K. Warfield, Suthirth Vaidya, Abhijith Chunduru, Ramanathan Muthuganapathy, Ganapathy Krishnamurthi, Andrew Jesson, Tal Arbel, Oskar Maier, Heinz Handels, Leonardo O. Iheme, Devrim Unay, Saurabh Jain, Diana M. Sima, Dirk Smeets, Mohsen Ghafoorian, Bram Platel, Ariel Birenbaum, Hayit Greenspan, Pierre Louis Bazin, Peter A. Calabresi, Ciprian M. Crainiceanu, Lotta M. Ellingsen, Daniel S. Reich, Jerry L. Prince, and Dzung L. Pham. 2017. "Longitudinal Multiple Sclerosis Lesion Segmentation: Resource and Challenge." *NeuroImage* 148:77–102.

Cardoso, M. Jorge, Kelvin Leung, Marc Modat, Shiva Keihaninejad, David Cash, Josephine Barnes, Nick C. Fox, and Sebastien Ourselin. 2013. "STEPS: Similarity and Truth Estimation for Propagated Segmentations and Its Application to Hippocampal Segmentation and Brain Parcelation." *Medical Image Analysis* 17(6):671–84.

Cardoso, M. Jorge, Marc Modat, Robin Wolz, Andrew Melbourne, David Cash, Daniel Rueckert, and Sebastien Ourselin. 2015. "Geodesic Information Flows: Spatially-Variant Graphs and Their Application to Segmentation and Fusion." *IEEE Transactions on Medical Imaging* 34(9):1976–88.

Caroli, Anna and Giovanni B. Frisoni. 2009. "Quantitative Evaluation of Alzheimer's Disease." *Expert Review of Medical Devices* 6(5):569–88.

Cedazo-Minguez, A, C Graff, G Johansson, L Jönsson, M Kivipelto, L O Tjernberg, Alzheimer Biology, Karolinska Institutet, Sweden ;. Huddinge, Bengt Winblad, Philippe Amouyel, Sandrine Andrieu, Clive Ballard, Carol Brayne, Henry Brodaty, Angel Cedazo-Minguez, Bruno Dubois, David Edvardsson, Howard Feldman, Laura Fratiglioni, Giovanni B. Frisoni, Serge Gauthier, Jean Georges, Caroline Graff, Khalid Iqbal, Frank Jessen, Gunilla Johansson, Linus Jönsson, Miia Kivipelto, Martin Knapp, Francesca

Mangialasche, René Melis, Agneta Nordberg, Marcel Olde Rikkert, Chengxuan Qiu, Thomas P. Sakmar, Philip Scheltens, Lon S. Schneider, Reisa Sperling, Lars O Tjernberg, Gunhild Waldemar, Anders Wimo, and Henrik Zetterberg. 2016. *The Lancet Neurology Commission Defeating Alzheimer's Disease and Other Dementias: A Priority for European Science and Society*. Vol. 15.

Chard, D. T., C. M. Griffin, G. J. M. Parker, R. Kapoor, A. J. Thompson, and D. H. Miller. 2002. "Brain Atrophy in Clinically Early Relapsing–Remitting Multiple Sclerosis." *Brain* 125(2):327–37.

Chard, Declan T, Jonathan S. Jackson, David H. Miller, and Claudia A. M. Wheeler-Kingshott. 2010. "Reducing the Impact of White Matter Lesions on Automated Measures of Brain Gray and White Matter Volumes." *Journal of Magnetic Resonance Imaging : JMRI* 32(1):223–28.

Chard, Declan T., Jonathan S. Jackson, David H. Miller, and Claudia A. M. Wheeler-Kingshott. 2010. "Reducing the Impact of White Matter Lesions on Automated Measures of Brain Gray and White Matter Volumes." *Journal of Magnetic Resonance Imaging* 32(1):223–28.

Chauhan, Richa Singh, Karthik Kulanthaivelu, Nihar Kathrani, Abhishek Kotwal, Maya Dattatraya Bhat, Jitender Saini, Chandrajit Prasad, Dhritiman Chakrabarti, Vani Santosh, Alok Mohan Uppar, and Dwarakanath Srinivas. 2021. "Prediction of H3K27M Mutation Status of Diffuse Midline Gliomas Using MRI Features." *Journal of Neuroimaging* jon.12905.

Chen, J. Y. and F. J. Lexa. 2017. "Baseline Survey of the Neuroradiology Work Environment in the United States with Reported Trends in Clinical Work, Nonclinical Work, Perceptions of Trainees, and Burnout Metrics." *American Journal of Neuroradiology* 38(7):1284–91.

Choi, Hyunmi, Randall L. Sell, Leslie Lenert, Peter Muennig, Robert R. Goodman, Frank G. Gilliam, and John B. Wong. 2008. "Epilepsy Surgery for Pharmacoresistant Temporal Lobe Epilepsy: A Decision Analysis." *JAMA - Journal of the American Medical Association* 300(21):2497–2505.

Clarke, Laurence P., Ram D. Sriram, and Linda Beth Schilling. 2008. "Imaging as a Biomarker: Standards for Change Measurements in Therapy Workshop Summary." *Academic Radiology* 15(4):501–30.

Clarkson, Matthew J., Gergely Zombori, Steve Thompson, Johannes Totz, Yi Song, Miklos Espak, Stian Johnsen, David Hawkes, and Sébastien Ourselin. 2015. "The NifTK Software Platform for Image-Guided Interventions: Platform Overview and NiftyLink Messaging." *International Journal of Computer Assisted Radiology and Surgery* 10(3):301–16.

CMSC. 2021. "Consortium of Multiple Sclerosis Centers (CMSC)." Retrieved August 23, 2021 (https://www.mscare.org/?).

Coan, A. C., B. Kubota, F. P. G. Bergo, B. M. Campos, and Fernando Cendes. 2014. "3T MRI Quantification of Hippocampal Volume and Signal in Mesial Temporal Lobe Epilepsy Improves Detection of Hippocampal Sclerosis." *American Journal of Neuroradiology* 35(1):77–83.

Cohen, J. 2013. *Statistical Power Analysis for the Behavioral Sciences*. Routledge.

Cohen, Jacob. 1960. “A Coefficient of Agreement for Nominal Scales.” *Educational and Psychological Measurement* 20(1):37–46.

Cole, J. H., S. J. Ritchie, M. E. Bastin, M. C. Valdés Hernández, S. Muñoz Maniega, N. Royle, J. Corley, A. Pattie, S. E. Harris, Q. Zhang, N. R. Wray, P. Redmond, R. E. Marioni, J. M. Starr, S. R. Cox, J. M. Wardlaw, D. J. Sharp, and I. J. Deary. 2018. “Brain Age Predicts Mortality.” *Molecular Psychiatry* 23(5):1385–92.

Cole, James H. and Katja Franke. 2017. “Predicting Age Using Neuroimaging: Innovative Brain Ageing Biomarkers.” *Trends in Neurosciences* 40(12):681–90.

Cole, James H., Robert Leech, and David J. Sharp. 2015. “Prediction of Brain Age Suggests Accelerated Atrophy after Traumatic Brain Injury.” *Annals of Neurology* 77(4):571–81.

Cole, James H., Joel Raffel, Tim Friede, Arman Eshaghi, Wallace J. Brownlee, Declan Chard, Nicola De Stefano, Christian Enzinger, Lukas Pirpamer, Massimo Filippi, Claudio Gasperini, Maria Assunta Rocca, Alex Rovira, Serena Ruggieri, Jaume Sastre-Garriga, Maria Laura Stromillo, Bernard M. J. Uitdehaag, Hugo Vrenken, Frederik Barkhof, Richard Nicholas, and Olga Ciccarelli. 2020. “Longitudinal Assessment of Multiple Sclerosis with the Brain-Age Paradigm.” *Annals of Neurology* 88(1):93–105.

Cole, JH, J. Raffel, T. Friede, A. Eshaghi, W. Brownlee, D. Chard, N. De Stefano, C. Enzinger, L. Pirpamer, M. Filippi, C. Gasperini, MA Rocca, A. Rovira, S. Ruggieri, J. Sastre-Garriga, ML Stromillo, BMJ Uitdehaag, H. Vrenken, F. Barkhof, R. Nicholas, O. Ciccarelli, and on behalf of the MAGNIMS study group. 2019. “Accelerated Brain Ageing and Disability in Multiple Sclerosis.” *BioRxiv* 584888.

Combinostics. 2021. “Combinostics.” Retrieved August 23, 2021 (https://www.combinostics.com/).

Cook, Laura, Helen Souris, and Jeremy Isaacs. 2019. “London Memory Services 2019 Audit Report.” Retrieved (https://www.england.nhs.uk/london/wp-content/uploads/sites/8/2019/11/FINAL-London-memory-service-audit-2019.pdf).

Cook, M. J., D. R. Fish, S. D. Shorvon, K. Straughan, and J. M. Stevens. 1992. “Hippocampal Volumetric and Morphometric Studies in Frontal and Temporal Lobe Epilepsy.” *Brain : A Journal of Neurology* 115 (Pt 4(4):1001–15.

Coras, R. and I. Blümcke. 2015. “Clinico-Pathological Subtypes of Hippocampal Sclerosis in Temporal Lobe Epilepsy and Their Differential Impact on Memory Impairment.” *Neuroscience* 309:153–61.

cortechs.ai. 2021. “Cortechs.Ai Radiology Applications.” Retrieved August 23, 2021 (https://www.cortechs.ai/).

Counts, Scott E., Milos D. Ikonomovic, Natosha Mercado, Irving E. Vega, and Elliott J. Mufson. 2016. “Biomarkers for the Early Detection and Progression of Alzheimer’s Disease.” *Neurotherapeutics 2016 14:1* 14(1):35–53.

Crutch, Sebastian J., Manja Lehmann, Jonathan M. Schott, Gil D. Rabinovici, Martin N. Rossor, and Nick C. Fox. 2012. “Posterior Cortical Atrophy.” *The Lancet. Neurology* 11(2):170–78.

Crutch, Sebastian J., Jonathan M. Schott, Gil D. Rabinovici, Melissa Murray, Julie S. Snowden, Wiesje M. van der Flier, Bradford C. Dickerson, Rik Vandenberghe, Samrah Ahmed, Thomas H. Bak, Bradley F. Boeve, Christopher Butler, Stefano F. Cappa, Mathieu Ceccaldi, Leonardo Cruz de Souza, Bruno Dubois, Olivier Felician, Douglas

Galasko, Jonathan Graff-Radford, Neill R. Graff-Radford, Patrick R. Hof, Pierre Krolak-Salmon, Manja Lehmann, Eloi Magnin, Mario F. Mendez, Peter J. Nestor, Chiadi U. Onyike, Victoria S. Pelak, Yolande Pijnenburg, Silvia Primativo, Martin N. Rossor, Natalie S. Ryan, Philip Scheltens, Timothy J. Shakespeare, Aida Suárez González, David F. Tang-Wai, Keir X. X. Yong, Maria Carrillo, Nick C. Fox, and on behalf of the Alzheimer's Association ISTAARTAtypical Alzheimer's Disease and Associated Syndromes Professional Interest Alzheimer's Association ISTAART Atypical Alzheimer's Disease and Associated Syndromes Professional Interest Area. 2017. "Consensus Classification of Posterior Cortical Atrophy." *Alzheimer's & Dementia : The Journal of the Alzheimer's Association* 13(8):870–84.

Cunningham, E. L., B. McGuinness, B. Herron, and A. P. Passmore. 2015. "Dementia." *The Ulster Medical Journal* 84(2):79–87.

Danelakis, Antonios, Theoharis Theoharis, and Dimitrios A. Verganelakis. 2018. "Survey of Automated Multiple Sclerosis Lesion Segmentation Techniques on Magnetic Resonance Imaging." *Computerized Medical Imaging and Graphics : The Official Journal of the Computerized Medical Imaging Society* 70:83–100.

Darlix, Amélie, Jérémy Deverdun, Nicolas Menjot de Champfleur, Florence Castan, Sonia Zouaoui, Valérie Rigau, Michel Fabbro, Yordanka Yordanova, Emmanuelle Le Bars, Luc Bauchet, Catherine Gozé, and Hugues Duffau. 2017. "IDH Mutation and 1p19q Codeletion Distinguish Two Radiological Patterns of Diffuse Low-Grade Gliomas." *Journal of Neuro-Oncology 2017 133:1* 133(1):37–45.

deSouza, Nandita M., Eric Achten, Angel Alberich-Bayarri, Fabian Bamberg, Ronald Boellaard, Olivier Clément, Laure Fournier, Ferdia Gallagher, Xavier Golay, Claus Peter Heussel, Edward F. Jackson, Rashindra Manniesing, Marius E. Mayerhofer, Emanuele Neri, James O'Connor, Kader Karli Oguz, Anders Persson, Marion Smits, Edwin J. R. van Beek, and Christoph J. Zech. 2019. "Validated Imaging Biomarkers as Decision-Making Tools in Clinical Trials and Routine Practice: Current Status and Recommendations from the EIBALL* Subcommittee of the European Society of Radiology (ESR)." *Insights into Imaging* 10(1).

Despotović, Ivana, Bart Goossens, and Wilfried Philips. 2015. "MRI Segmentation of the Human Brain: Challenges, Methods, and Applications." *Computational and Mathematical Methods in Medicine* 2015:1–23.

DeToledo-Morrell, L., T. .. Stoub, M. Bulgakova, R. .. Wilson, D. .. Bennett, S. Leurgans, J. Wuu, and D. .. Turner. 2004. "MRI-Derived Entorhinal Volume Is a Good Predictor of Conversion from MCI to AD." *Neurobiology of Aging* 25(9):1197–1203.

Devanand, Davangere P., Xinhua Liu, Matthias H. Tabert, Gnanavalli Pradhaban, Katrina Cuasay, Karen Bell, Mony J. de Leon, Richard L. Doty, Yaakov Stern, and Gregory H. Pelton. 2008. "Combining Early Markers Strongly Predicts Conversion from Mild Cognitive Impairment to Alzheimer's Disease." *Biological Psychiatry* 64(10):871–79.

Drummond, Michael F., J. Sanford Schwartz, Bengt Jönsson, Bryan R. Luce, Peter J. Neumann, Uwe Siebert, and Sean D. Sullivan. 2008. "Key Principles for the Improved Conduct of Health Technology Assessments for Resource Allocation Decisions." *International Journal of Technology Assessment in Health Care* 24(03):244–58.

Duan, Yiran, Yicong Lin, Dennis Rosen, Jialin Du, Liu He, and Yuping Wang. 2020. "Identifying Morphological Patterns of Hippocampal Atrophy in Patients With Mesial Temporal Lobe Epilepsy and Alzheimer Disease." *Frontiers in Neurology* 11:21.

Dubois, Bruno, Howard H. Feldman, Claudia Jacova, Steven T. Dekosky, Pascale Barberger-Gateau, Jeff Rey Cummings, André Delacourte, Douglas Galasko, Serge Gauthier, Gregory Jicha, Kenichi Meguro, John O'brien, Florence Pasquier, Philippe Robert, Martin Rossor, Steven Salloway, Yaakov Stern, Pieter J. Visser, Philip Scheltens, Jeffrey Cummings, André Delacourte, Douglas Galasko, Serge Gauthier, Gregory Jicha, Kenichi Meguro, John O'brien, Florence Pasquier, Philippe Robert, Martin Rossor, Steven Salloway, Yaakov Stern, Pieter J. Visser, and Philip Scheltens. 2007. *Research Criteria for the Diagnosis of Alzheimer's Disease: Revising the NINCDS-ADRDA Criteria*. Vol. 6. Elsevier.

Duncan, J. S. and H. J. Sagar. 1987. "Seizure Characteristics, Pathology, and Outcome after Temporal Lobectomy." *Neurology* 37(3):405–9.

Durlak, J. A. 2009. "How to Select, Calculate, and Interpret Effect Sizes." *Journal of Pediatric Psychology* 34(9):917–28.

Dwyer, Michael G., Niels Bergsland, Deepa P. Ramasamy, Dejan Jakimovski, Bianca Weinstock-Guttman, and Robert Zivadinov. 2018. "Atrophied Brain Lesion Volume: A New Imaging Biomarker in Multiple Sclerosis." *Journal of Neuroimaging* 28(5):490–95.

Dwyer, Michael G., Niels Bergsland, Deepa P. Ramasamy, Bianca Weinstock-Guttman, Michael H. Barnett, Chenyu Wang, Davorka Tomic, Diego Silva, and Robert Zivadinov. 2019. "Salient Central Lesion Volume: A Standardized Novel Fully Automated Proxy for Brain FLAIR Lesion Volume in Multiple Sclerosis." *Journal of Neuroimaging* 29(5):615–23.

Eidel, Oliver, Sina Burth, Jan-Oliver Neumann, Pascal J. Kieslich, Felix Sahm, Christine Jungk, Philipp Kickingereder, Sebastian Bickelhaupt, Sibu Mundiyanapurath, Philipp Bäumer, Wolfgang Wick, Heinz-Peter Schlemmer, Karl Kiening, Andreas Unterberg, Martin Bendszus, and Alexander Radbruch. 2017. "Tumor Infiltration in Enhancing and Non-Enhancing Parts of Glioblastoma: A Correlation with Histopathology" edited by C. Kleinschnitz. *PLOS ONE* 12(1):e0169292.

Ellingson, BM, M. Bendszus, J. Boxerman, D. Barboriak, BJ Erickson, M. Smits, SJ Nelson, E. Gerstner, B. Alexander, G. Goldmacher, W. Wick, M. Vogelbaum, M. Weller, E. Galanis, J. Kalpathy-Cramer, L. Shankar, P. Jacobs, WB Pope, D. Yang, C. Chung, MV Knopp, S. Cha, MJ van den Bent, S. Chang, WK Yung, TF Cloughesy, PY Wen, and MR Gilbert. 2015. "Consensus Recommendations for a Standardized Brain Tumor Imaging Protocol in Clinical Trials." *Neuro-Oncology* 17(9):1188–98.

Engel J., Jr. 2001. "Mesial Temporal Lobe Epilepsy: What Have We Learned?" *Neuroscientist* 7(4):340–52.

Engel, J., S. Wiebe, J. French, M. Sperling, P. Williamson, D. Spencer, R. Gumnit, C. Zahn, E. Westbrook, and B. Enos. 2003. "Practice Parameter: Temporal Lobe and Localized Neocortical Resections for Epilepsy." *Neurology* 60(4):538–47.

EPAD. 2021. "EPAD | European Prevention of Alzheimer's Dementia Consortium." Retrieved August 23, 2021 (http://ep-ad.org/).

Eshaghi, Arman, Ferran Prados, Wallace J. Brownlee, Daniel R. Altmann, Carmen Tur, M. Jorge Cardoso, Floriana De Angelis, Steven H. van de Pavert, Niamh Cawley, Nicola De Stefano, M. Laura Stromillo, Marco Battaglini, Serena Ruggieri, Claudio Gasperini, Massimo Filippi, Maria A. Rocca, Alex Rovira, Jaume Sastre-Garriga, Hugo Vrenken, Cyra E. Leurs, Joep Killestein, Lukas Pirpamer, Christian Enzinger, Sebastien Ourselin, Claudia A. M. Gandin. Wheeler-Kingshott, Declan Chard, Alan J. Thompson, Daniel C.

Alexander, Frederik Barkhof, and Olga Ciccarelli. 2018. “Deep Gray Matter Volume Loss Drives Disability Worsening in Multiple Sclerosis.” *Annals of Neurology* 83(2).

Eshaghi, Arman, Alexandra L. Young, Peter A. Wijeratne, Ferran Prados, Douglas L. Arnold, Sridar Narayanan, Charles R. G. Guttmann, Frederik Barkhof, Daniel C. Alexander, Alan J. Thompson, Declan Chard, and Olga Ciccarelli. 2021. “Identifying Multiple Sclerosis Subtypes Using Unsupervised Machine Learning and MRI Data.” *Nature Communications* 12(1):1–12.

ESR. 2020. “ESR Statement on the Validation of Imaging Biomarkers.” *Insi* 11(76).

ESR. 2021a. “AI | European Society of Radiology.” Retrieved August 23, 2021 (https://www.myesr.org/ai).

ESR. 2021b. “European Imaging Biomarkers Alliance - EIBALL | European Society of Radiology.” Retrieved August 23, 2021 (https://www.myesr.org/research/european-imaging-biomarkers-alliance-eiball).

EU. 2021. “Overview | Public Health.” Retrieved August 23, 2021 (https://ec.europa.eu/health/md_sector/overview_en).

Farid, Nikdokht, Holly M. Girard, Nobuko Kemmotsu, Michael E. Smith, Sebastian W. Magda, Wei Y. Lim, Roland R. Lee, and Carrie R. Mcdonald. 2012. “Temporal Lobe Epilepsy: Quantitative MR Volumetry in Detection of Hippocampal Atrophy 1.” *Radiology.Rsna.Org n Radiology* 264(2).

FDA. 2021. “Software as a Medical Device (SaMD) Action Plan.”

Feldman, Mitchell D., Amy J. Petersen, Leah S. Karliner, and Jeffrey A. Tice. 2008. “Who Is Responsible for Evaluating the Safety and Effectiveness of Medical Devices? The Role of Independent Technology Assessment.” *Journal of General Internal Medicine* 23 Suppl 1(Suppl 1):57–63.

Filippi, Massimo, Paolo Preziosa, Brenda L. Banwell, Frederik Barkhof, Olga Ciccarelli, Nicola De Stefano, Jeroen J. G. G. Geurts, Friedemann Paul, Daniel S. Reich, Ahmed T. Toosy, Anthony Traboulsee, Mike P. Wattjes, Tarek A. Yousry, Achim Gass, Catherine Lubetzki, Brian G. Weinshenker, and Maria A. Rocca. 2019. *Assessment of Lesions on Magnetic Resonance Imaging in Multiple Sclerosis: Practical Guidelines*. Vol. 142. Oxford University Press.

Fischl, Bruce. 2012. “FreeSurfer.” *NeuroImage* 62(2):774–81.

Fischl, Bruce, David H. Salat, Evelina Busa, Marilyn Albert, Megan Dieterich, Christian Haselgrove, Andre Van Der Kouwe, Ron Killiany, David Kennedy, Shuna Klaveness, Albert Montillo, Nikos Makris, Bruce Rosen, and Anders M. Dale. 2002. “Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain.” *Neuron* 33(3):341–55.

Fischl, Bruce, Martin I. Sereno, and Anders M. Dale. 1999. “Cortical Surface-Based Analysis.” *NeuroImage* 9(2):195–207.

Flanders, Adam E. and Paras Lakhani. 2012. “Radiology Reporting and Communications. A Look Forward.” *Neuroimaging Clinics of North America* 22(3):477–96.

Fortin, Jean-Philippe, Nicholas Cullen, Yvette I. Sheline, Warren D. Taylor, Irem Aselcioglu, Philip A. Cook, Phil Adams, Crystal Cooper, Maurizio Fava, Patrick J. McGrath, Melvin McInnis, Mary L. Phillips, Madhukar H. Trivedi, Myrna M. Weissman, and Russell T.

Shinohara. 2018. “Harmonization of Cortical Thickness Measurements across Scanners and Sites.” *NeuroImage* 167:104–20.

Fouke, Sarah Jost, Tammie Benzinger, Daniel Gibson, Timothy C. Ryken, Steven N. Kalkanis, and Jeffrey J. Olson. 2015. “The Role of Imaging in the Management of Adults with Diffuse Low Grade Glioma.” *Journal of Neuro-Oncology* 125(3):457–79.

Franke, Katja and Christian Gaser. 2012. “Longitudinal Changes in Individual BrainAGE in Healthy Aging, Mild Cognitive Impairment, and Alzheimer’s Disease.” *GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry* 25(4):235–45.

Freeborough, Peter A. and Nick C. Fox. 1997. “The Boundary Shift Integral: An Accurate and Robust Measure of Cerebral Volume Changes from Registered Repeat MRI.” *IEEE Transactions on Medical Imaging* 16(5):623–29.

Frisoni, Giovanni B., Marina Boccardi, Frederik Barkhof, Kaj Blennow, Stefano Cappa, Konstantinos Chiotis, Jean-Francois Démonet, Valentina Garibotto, Panteleimon Giannakopoulos, Anton Gietl, Oskar Hansson, Karl Herholz, Clifford R. Jack, Flavio Nobili, Agneta Nordberg, Heather M. Snyder, Mara Ten Kate, Andrea Varrone, Emiliano Albanese, Stefanie Becker, Patrick Bossuyt, Maria C. Carrillo, Chiara Cerami, Bruno Dubois, Valentina Gallo, Ezio Giacobini, Gabriel Gold, Samia Hurst, Anders Lönneborg, Karl-Olof Lovblad, Niklas Mattsson, José-Luis Molinuevo, Andreas U. Monsch, Urs Mosimann, Alessandro Padovani, Agnese Picco, Corinna Porteri, Osman Ratib, Laure Saint-Aubert, Charles Scerri, Philip Scheltens, Jonathan M. Schott, Ida Sonni, Stefan Teipel, Paolo Vineis, Pieter Jelle Visser, Yutaka Yasui, and Bengt Winblad. 2017. “Strategic Roadmap for an Early Diagnosis of Alzheimer’s Disease Based on Biomarkers.” *The Lancet Neurology* 16(8):661–76.

Gafson, Arie, Matt J. Craner, and Paul M. Matthews. 2017. “Personalised Medicine for Multiple Sclerosis Care.” *Multiple Sclerosis* 23(3):362–69.

Galanis, Evanthia, Jan C. Buckner, Matthew J. Maurer, Rene Sykora, René Castillo, Karla V. Ballman, and Bradley J. Erickson. 2006. “Validation of Neuroradiologic Response Assessment in Gliomas: Measurement by RECIST, Two-Dimensional, Computer-Assisted Tumor Area, and Computer-Assisted Tumor Volume Methods1.” *Neuro-Oncology* 8(2):156–65.

Ganeshan, Dhakshinamoorthy, Phuong-Anh Thi Duong, Linda Probyn, Leon Lenchik, Tatum A. McArthur, Michele Retrouvey, Emily H. Ghobadi, Stephane L. Desouches, David Pastel, and Isaac R. Francis. 2018. “Structured Reporting in Radiology.” *Academic Radiology* 25(1):66–73.

George, Allan, Ruben Kuzniecky, Henry Rusinek, and Heath R. Pardoe. 2019. “Standardized Brain MRI Acquisition Protocols Improve Statistical Power in Multicenter Quantitative Morphometry Studies.” *Journal of Neuroimaging*.

Geurts, J. J. G., S. D. Roosendaal, M. Calabrese, O. Ciccarelli, F. Agosta, D. T. Chard, A. Gass, E. Huerga, B. Moraal, D. Pareto, M. A. Rocca, M. P. Wattjes, T. A. Yousry, B. M. J. Uitdehaag, and F. Barkhof. 2011. “Consensus Recommendations for MS Cortical Lesion Scoring Using Double Inversion Recovery MRI.” *Neurology* 76(5):418–24.

Giovannoni, Gavin, Helmut Butzkueven, Suhayl Dhib-Jalbut, Jeremy Hobart, Gisela Kobelt, George Pepper, Maria Pia Sormani, Christoph Thalheim, Anthony Traboulsee, and Timothy Vollmer. 2016. “Brain Health: Time Matters in Multiple Sclerosis.” *Multiple Sclerosis and Related Disorders* 9:S5–48.

Good, Catriona D., Ingrid S. Johnsrude, John Ashburner, Richard N. A. Henson, Karl J. Friston, and Richard S. J. Frackowiak. 2001. “A Voxel-Based Morphometric Study of Ageing in 465 Normal Adult Human Brains.” *NeuroImage* 14(1):21–36.

Greenhalgh, Trisha, Joseph Wherton, Chrysanthi Papoutsi, Jennifer Lynch, Gemma Hughes, Christine A’Court, Susan Hinder, Nick Fahy, Rob Procter, and Sara Shaw. 2017. “Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies.” *Journal of Medical Internet Research* 19(11):e367.

Gutman, David A., Lee A. D. Cooper, Scott N. Hwang, Chad A. Holder, JingJing Gao, Tarun D. Aurora, William D. Dunn, Lisa Scarpace, Tom Mikkelsen, Rajan Jain, Max Wintermark, Manal Jilwan, Prashant Raghavan, Erich Huang, Robert J. Clifford, Pattanasak Mongkolwat, Vladimir Kleper, John Freymann, Justin Kirby, Pascal O. Zinn, Carlos S. Moreno, Carl Jaffe, Rivka Colen, Daniel L. Rubin, Joel Saltz, Adam Flanders, and Daniel J. Brat. 2013. “MR Imaging Predictors of Molecular Profile and Survival: Multi-Institutional Study of the TCGA Glioblastoma Data Set.” *Radiology* 267(2):560–69.

Hagens, Marloes H. J., Jessica Burggraaff, Iris D. Kilsdonk, Marlieke L. de Vos, Niamh Cawley, Emilia Sbardella, Michaela Andelova, Michael Amann, Johanna M. Lieb, Patrizia Pantano, Birgit I. Lissenberg-Witte, Joep Killestein, Celia Oreja-Guevara, Olga Ciccarelli, Claudio Gasperini, Carsten Lukas, Mike P. Wattjes, and Frederik Barkhof. 2018. “Three-Tesla MRI Does Not Improve the Diagnosis of Multiple Sclerosis.” *Neurology* 91(3):e249–57.

Haider, Lukas, Ferran Prados, Karen Chung, Olivia Goodkin, Baris Kanber, Carole Sudre, Marios Yiannakas, Rebecca S. Samson, Stephanie Mangesius, Alan J. Thompson, Claudia A. M. Gandini Wheeler-Kingshott, Olga Ciccarelli, Declan T. Chard, and Frederik Barkhof. 2021. “Cortical Involvement Determines Impairment 30 Years after a Clinically Isolated Syndrome.” *Brain* 144(5).

Haller, Sven, Pavel Falkovskiy, Reto Meuli, Jean-Philippe Thiran, Gunnar Krueger, Karl-Olof Lovblad, Tobias Kober, Alexis Roche, and Bénédicte Marechal. 2016. “Basic MR Sequence Parameters Systematically Bias Automated Brain Volume Estimation.” *Neuroradiology* 58(11):1153–60.

Han, Xiao, Jorge Jovicich, David Salat, Andre van der Kouwe, Brian Quinn, Silvester Czanner, Evelina Busa, Jenni Pacheco, Marilyn Albert, Ronald Killiany, Paul Maguire, Diana Rosas, Nikos Makris, Anders Dale, Bradford Dickerson, and Bruce Fischl. 2006. “Reliability of MRI-Derived Measurements of Human Cerebral Cortical Thickness: The Effects of Field Strength, Scanner Upgrade and Manufacturer.” *NeuroImage* 32(1):180–94.

Harper, Lorna, Frederik Barkhof, Philip Scheltens, Jonathan M. Schott, and Nick C. Fox. 2014. “An Algorithmic Approach to Structural Imaging in Dementia.” *Journal of Neurology, Neurosurgery, and Psychiatry* 85(6):692–98.

Harper, Lorna, Giorgio G. Fumagalli, Frederik Barkhof, Philip Scheltens, John T. O’Brien, Femke Bouwman, Emma J. Burton, Jonathan D. Rohrer, Nick C. Fox, Gerard R. Ridgway, and Jonathan M. Schott. 2016. “MRI Visual Rating Scales in the Diagnosis of Dementia: Evaluation in 184 Post-Mortem Confirmed Cases.” *Brain* 139(4):1211–25.

Heckemann, Rolf A., Joseph V. Hajnal, Paul Aljabar, Daniel Rueckert, and Alexander Hammers. 2006. “Automatic Anatomical Brain MRI Segmentation Combining Label Propagation and Decision Fusion.” *NeuroImage* 33(1):115–26.

Heckemann, Rolf A., Alexander Hammers, Daniel Rueckert, Richard I. Aviv, Christopher J. Harvey, and Joseph V Hajnal. 2008. “Automatic Volumetry on MR Brain Images Can Support Diagnostic Decision Making.” *BMC Medical Imaging* 8(1):9.

Hedderich, Dennis M., Michael Dieckmeyer, Tiberiu Andrisan, Marion Ortner, Lioba Grundl, Simon Schön, Per Suppa, Tom Finck, Kornelia Kreiser, Claus Zimmer, Igor Yakushev, and Timo Grimmer. 2020. “Normative Brain Volume Reports May Improve Differential Diagnosis of Dementing Neurodegenerative Diseases in Clinical Practice.” *European Radiology* 30(5):2821–29.

Hedderich, Dennis M., Judith E. Spiro, Oliver Goldhardt, Johannes Kaesmacher, Benedikt Wiestler, Igor Yakushev, Claus Zimmer, Tobias Boeckh-Behrens, and Timo Grimmer. 2018. “Increasing Diagnostic Accuracy of Mild Cognitive Impairment Due to Alzheimer’s Disease by User-Independent, Web-Based Whole-Brain Volumetry.” *Journal of Alzheimer’s Disease* 65(4):1459–67.

Hennessy, M. J., R. D. Elwes, C. D. Binnie, and C. E. Polkey. 2000. “Failed Surgery for Epilepsy: A Study of Persistence and Recurrence of Seizures Following Temporal Resection.” *Brain* 123(12):2445–66.

Hirai, Toshinori, Yukunori Korogi, Kazuhiro Yoshizumi, Yoshinori Shigematsu, Takeshi Sugahara, and Mutsumasa Takahashi. 2000. “Limbic Lobe of the Human Brain: Evaluation with Turbo Fluid-Attenuated Inversion-Recovery MR Imaging.” *Radiology* 215(2):470–75.

Hosny, Ahmed, Chintan Parmar, John Quackenbush, Lawrence H. Schwartz, and Hugo J. W. L. Aerts. 2018. “Artificial Intelligence in Radiology.” *Nature Reviews Cancer* 18(8):500–510.

Hu, Wen Han, Li Na Liu, Bao Tian Zhao, Xiu Wang, Chao Zhang, Xiao Qiu Shao, Kai Zhang, Yan Shan Ma, Lin Ai, Jun Ju Li, and Jian Guo Zhang. 2018. “Use of an Automated Quantitative Analysis of Hippocampal Volume, Signal, and Glucose Metabolism to Detect Hippocampal Sclerosis.” *Frontiers in Neurology* 9(OCT).

Huisman, Merel, Erik Ranschaert, William Parker, Domenico Mastrodicasa, Martin Koci, Daniel Pinto de Santos, Francesca Coppola, Sergey Morozov, Marc Zins, Cedric Bohyn, Ural Koç, Jie Wu, Satyam Veean, Dominik Fleischmann, Tim Leiner, and Martin J. Willemink. 2021. “An International Survey on AI in Radiology in 1,041 Radiologists and Radiology Residents Part 1: Fear of Replacement, Knowledge, and Attitude.” *European Radiology* 1–9.

Huppertz, Hans Jürgen, Jan Wagner, Bernd Weber, Patrick House, and Horst Urbach. 2011. “Automated Quantitative FLAIR Analysis in Hippocampal Sclerosis.” *Epilepsy Research* 97(1–2):146–56.

Hyare, Harpreet, Rice L, Thust S, Nachev P, Jha A, Milic M, Brandner S, Rees J, Louise Rice, Stefanie Thust, Parashkev Nachev, Ashwani Jha, Marina Milic, Sebastian Brandner, and Jeremy Rees. 2019. “Modelling MR and Clinical Features in Grade II/III Astrocytomas to Predict IDH Mutation Status.” 114:120–27.

icometrix. 2021. “Transforming Patient Care through Imaging AI | Icometrix.” Retrieved August 23, 2021 (https://www.icometrix.com/).

Jack, C. R., D. W. Dickson, J. E. Parisi, Y. C. Xu, R. H. Cha, P. C. O’Brien, S. D. Edland, G. E. Smith, B. F. Boeve, E. G. Tangalos, E. Kokmen, and R. C. Petersen. 2002. “Antemortem MRI Findings Correlate with Hippocampal Neuropathology in Typical Aging and

Dementia." *Neurology* 58(5):750–57.

Jack, Clifford R., Josephine Barnes, Matt A. Bernstein, Bret J. Borowski, James Brewer, Shona Clegg, Anders M. Dale, Owen Carmichael, Christopher Ching, Charles DeCarli, Rahul S. Desikan, Christine Fennema-Notestine, Anders M. Fjell, Evan Fletcher, Nick C. Fox, Jeff Gunter, Boris A. Gutman, Dominic Holland, Xue Hua, Philip Insel, Kejal Kantarci, Ron J. Killiany, Gunnar Krueger, Kelvin K. Leung, Scott Mackin, Pauline Maillard, Ian B. Malone, Niklas Mattsson, Linda McEvoy, Marc Modat, Susanne Mueller, Rachel Nosheny, Sebastien Ourselin, Norbert Schuff, Matthew L. Senjem, Alix Simonson, Paul M. Thompson, Dan Rettmann, Prashanthi Vemuri, Kristine Walhovd, Yansong Zhao, Samantha Zuk, and Michael Weiner. 2015. "Magnetic Resonance Imaging in Alzheimer's Disease Neuroimaging Initiative 2." *Alzheimer's and Dementia* 11(7):740–56.

Jack, Clifford R., Matt A. Bernstein, Nick C. Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J. Britson, Jennifer L. Whitwell, Chadwick Ward, Anders M. Dale, Joel P. Felmlee, Jeffrey L. Gunter, Derek L. G. Hill, Ron Killiany, Norbert Schuff, Sabrina Fox-Bosetti, Chen Lin, Colin Studholme, Charles S. DeCarli, Gunnar Krueger, Heidi A. Ward, Gregory J. Metzger, Katherine T. Scott, Richard Mallozzi, Daniel Blezek, Joshua Levy, Josef P. Debbins, Adam S. Fleisher, Marilyn Albert, Robert Green, George Bartzokis, Gary Glover, John Mugler, and Michael W. Weiner. 2008. "The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI Methods." *Journal of Magnetic Resonance Imaging* 27(4):685–91.

Jackson, G. D., A. Connelly, J. S. Duncan, R. A. Grünewald, and D. G. Gadian. 1993. "Detection of Hippocampal Pathology in Intractable Partial Epilepsy: Increased Sensitivity with Quantitative Magnetic Resonance T2 Relaxometry." *Neurology* 43(9):1793–99.

Jain, Saurabh, Diana M. Sima, Annemie Ribbens, Melissa Cambron, Anke Maertens, Wim Van Hecke, Johan De Mey, Frederik Barkhof, Martijn D. Steenwijk, Marita Daams, Frederik Maes, Sabine Van Huffel, Hugo Vrenken, and Dirk Smeets. 2015. "Automatic Segmentation and Volumetry of Multiple Sclerosis Brain Lesions from MR Images." *NeuroImage: Clinical* 8:367–75.

Jordan, J. E. and Adam E. Flanders. 2019. "The ASNR-ACR-RSNA Common Data Elements Project: What Will It Do for the House of Neuroradiology?" *American Journal of Neuroradiology* 40(1):14–18.

jung diagnostics. 2021. "Jung Diagnostics." Retrieved August 23, 2021 (https://www.jung-diagnostics.de/).

Kappos, Ludwig, Mark S. Freedman, Chris H. Polman, Gilles Edan, Hans Peter Hartung, David H. Miller, Xavier Montalbán, Frederik Barkhof, Ernst Wilhelm Radü, Lars Bauer, Susanne Dahms, Vivian Lanius, Christoph Pohl, and Rupert Sandbrink. 2007. "Effect of Early versus Delayed Interferon Beta-1b Treatment on Disability after a First Clinical Event Suggestive of Multiple Sclerosis: A 3-Year Follow-up Analysis of the BENEFIT Study." *Lancet* 370(9585):389–97.

ten Kate, Mara, Frederik Barkhof, Marina Boccardi, Pieter Jelle Visser, Clifford R. Jack, Karl-Olof Lovblad, Giovanni B. Frisoni, Philip Scheltens, and Geneva Task Force for the Roadmap of Alzheimer's Biomarkers. 2017. "Clinical Validity of Medial Temporal Atrophy as a Biomarker for Alzheimer's Disease in the Context of a Structured 5-Phase Development Framework." *Neurobiology of Aging* 52:167-182.e1.

Kessler, Larry G., Huiman X. Barnhart, Andrew J. Buckler, Kingshuk Roy Choudhury, Marina V. Kondratovich, Alicia Toledano, Alexander R. Guimaraes, Ross Filice, Zheng Zhang, and Daniel C. Sullivan. 2015. “The Emerging Science of Quantitative Imaging Biomarkers Terminology and Definitions for Scientific Studies and Regulatory Submissions.” *Statistical Methods in Medical Research* 24(1):9–26.

Kim, Hosung, Marie Chupin, Olivier Colliot, Boris C. Bernhardt, Neda Bernasconi, and Andrea Bernasconi. 2012. “Automatic Hippocampal Segmentation in Temporal Lobe Epilepsy: Impact of Developmental Abnormalities.” *NeuroImage* 59(4):3178–86.

Koedam, Esther L. G. E., Manja Lehmann, Wiesje M. van der Flier, Philip Scheltens, Yolande A. L. Pijnenburg, Nick Fox, Frederik Barkhof, and Mike P. Wattjes. 2011. “Visual Assessment of Posterior Atrophy Development of a MRI Rating Scale.” *European Radiology* 21(12):2618–25.

Kommers, Ivar, David Bouget, André Pedersen, Roelant S. Eijgelaar, Hilko Ardon, Frederik Barkhof, Lorenzo Bello, Mitchel S. Berger, Marco Conti Nibali, Julia Furtner, Even H. Fyllingen, Shawn Hervey-Jumper, Albert J. S. Idema, Barbara Kiesel, Alfred Kloet, Emmanuel Mandonnet, Domenique M. J. Müller, Pierre A. Robe, Marco Rossi, Lisa M. Sagberg, Tommaso Sciortino, Wimar A. van den Brink, Michiel Wagemakers, Georg Widhalm, Marnix G. Witte, Aeilko H. Zwinderman, Ingerid Reinertsen, Ole Solheim, and Philip C. De Witt Hamer. 2021. “Glioblastoma Surgery Imaging—Reporting and Data System: Standardized Reporting of Tumor Volume, Location, and Resectability Based on Automated Segmentations.” *Cancers* 13(12):2854.

Konrad, C., T. Ukas, C. Nebel, V. Arolt, A. W. Toga, and K. L. Narr. 2009. “Defining the Human Hippocampus in Cerebral Magnetic Resonance Images—An Overview of Current Segmentation Protocols.” *NeuroImage* 47(4):1185–95.

Koutsouleris, Nikolaos, Christos Davatzikos, Stefan Borgwardt, Christian Gaser, Ronald Bottlender, Thomas Frodl, Peter Falkai, Anita Riecher-Rössler, Hans Jürgen Möller, Maximilian Reiser, Christos Pantelis, and Eva Meisenzahl. 2014. “Accelerated Brain Aging in Schizophrenia and beyond: A Neuroanatomical Marker of Psychiatric Disorders.” *Schizophrenia Bulletin* 40(5):1140–53.

Kuhle, J., G. Disanto, R. Dobson, R. Adiutori, L. Bianchi, J. Topping, JP Bestwick, U. C. Meier, M. Marta, G. Dalla Costa, T. Runia, E. Evdoshenko, N. Lazareva, E. Thouvenot, P. Iaffaldano, V. Direnzo, M. Khademi, F. Piehl, M. Comabella, M. Sombekke, J. Killestein, H. Hegen, S. Rauch, S. D’Alfonso, JC Alvarez-Cermeño, P. Kleinová, D. Horáková, R. Roesler, F. Lauda, S. Llufriu, T. Avsar, U. Uygunoglu, A. Altintas, S. Saip, T. Menge, C. Rajda, R. Bergamaschi, N. Moll, M. Khalil, R. Marignier, I. Dujmovic, H. Larsson, C. Malmestrom, E. Scarpini, C. Fenoglio, S. Wergeland, A. Laroni, V. Annibali, S. Romano, AD Martínez, A. Carra, M. Salvetti, A. Uccelli, Ø. Torkildsen, KM Myhr, D. Galimberti, K. Rejdak, J. Lycke, JL Frederiksen, J. Drulovic, C. Confavreux, D. Brassat, C. Enzinger, S. Fuchs, I. Bosca, J. Pelletier, C. Picard, E. Colombo, D. Franciotta, T. Derfuss, RLP Lindberg, Ö. Yaldizli, L. Vécsei, BC Kieseier, HP Hartung, P. Villoslada, A. Siva, A. Saiz, H. Tumani, E. Havrdová, LM Villar, M. Leone, N. Barizzone, F. Deisenhammer, C. Teunissen, X. Montalban, M. Tintoré, T. Olsson, M. Trojano, S. Lehmann, G. Castelnovo, S. Lapin, R. Hintzen, L. Kappos, R. Furlan, V. Martinelli, G. Comi, SV Ramagopalan, and G. Giovannoni. 2015. “Conversion from Clinically Isolated Syndrome to Multiple Sclerosis: A Large Multicentre Study.” *Multiple Sclerosis Journal* 21(8):1013–24.

Lakens, Daniël. 2013. “Calculating and Reporting Effect Sizes to Facilitate Cumulative

Science: A Practical Primer for t-Tests and ANOVAs." *Frontiers in Psychology* 4(NOV):863.

Lambert, Christian, Janakan Sam Narean, Philip Benjamin, Eva Zeestraten, Thomas R. Barrick, and Hugh S. Markus. 2015. "Characterising the Grey Matter Correlates of Leukoaraiosis in Cerebral Small Vessel Disease." *NeuroImage: Clinical* 9:194–205.

Lasocki, Arian, Mustafa Anjari, Suna Örs Kokurcan, and Stefanie C. Thust. 2020. "Conventional MRI Features of Adult Diffuse Glioma Molecular Subtypes: A Systematic Review." *Neuroradiology* 63(3):353–62.

Lasocki, Arian, Alpha Tsui, Frank Gaillard, Mark Tacey, Katharine Drummond, and Stephen Stuckey. 2017. "Reliability of Noncontrast-Enhancing Tumor as a Biomarker of IDH1 Mutation Status in Glioblastoma." *Journal of Clinical Neuroscience* 39:170–75.

van Leeuwen, Kicky G., Steven Schalekamp, Matthieu J. C. M. Rutten, Bram van Ginneken, and Maarten de Rooij. 2021. "Artificial Intelligence in Radiology: 100 Commercially Available Products and Their Scientific Evidence." *European Radiology* 1–8.

Lencz, Todd, Gregory McCarthy, Richard A. Bronen, Tammy M. Scott, Jaime A. Inserni, Kimberlee J. Sass, Robert A. Novelly, Jung H. Kim, and Dennis D. Spencer. 1992. "Quantitative Magnetic Resonance Imaging in Temporal Lobe Epilepsy: Relationship to Neuropathology and Neuropsychological Function." *Annals of Neurology* 31(6):629–37.

Lin, E., DK Powell, and NJ Kagetsu. 2014. "Efficacy of a Checklist-Style Structured Radiology Reporting Template in Reducing Resident Misses on Cervical Spine Computed Tomography Examinations." *Journal of Digital Imaging* 27(5):588–93.

Lindig, Tobias, Raviteja Kotikalapudi, Daniel Schweikardt, Pascal Martin, Friedemann Bender, Uwe Klose, Ulrike Ernemann, Niels K. Focke, and Benjamin Bender. 2018. "Evaluation of Multimodal Segmentation Based on 3D T1-, T2- and FLAIR-Weighted Images – the Difficulty of Choosing." *NeuroImage* 170:210–21.

Losseff, N. A., L. Wang, H. M. Lai, D. S. Yoo, M. L. Gawne-Cain, W. I. McDonald, D. H. Miller, and A. J. Thompson. 1996. "Progressive Cerebral Atrophy in Multiple Sclerosis A Serial MRI Study." *Brain* 119(6):2009–19.

Louis, David N., Arie Perry, Guido Reifenberger, Andreas von Deimling, Dominique Figarella-Branger, Webster K. Cavenee, Hiroko Ohgaki, Otmar D. Wiestler, Paul Kleihues, and David W. Ellison. 2016. "The 2016 World Health Organization Classification of Tumors of the Central Nervous System: A Summary." *Acta Neuropathologica* 131(6):803–20.

Louis, S., M. Morita-Sherman, S. Jones, D. Vegh, W. Bingaman, I. Blumcke, N. Obuchowski, F. Cendes, and L. Jehi. 2020. "Hippocampal Sclerosis Detection with NeuroQuant Compared with Neuroradiologists." *American Journal of Neuroradiology*.

Lublin, Fred D., Stephen C. Reingold, Jeffrey A. Cohen, Gary R. Cutter, Per Soelberg Sørensen, Alan J. Thompson, Jerry S. Wolinsky, Laura J. Balcer, Brenda Banwell, Frederik Barkhof, Bruce Bebo, Peter A. Calabresi, Michel Clanet, Giancarlo Comi, Robert J. Fox, Mark S. Freedman, Andrew D. Goodman, Matilde Inglese, Ludwig Kappos, Bernd C. Kieseier, John A. Lincoln, Catherine Lubetzki, Aaron E. Miller, Xavier Montalban, Paul W. O'Connor, John Petkau, Carlo Pozzilli, Richard A. Rudick, Maria Pia Sormani, Olaf Stüve, Emmanuelle Waubant, and Chris H. Polman. 2014. "Defining the Clinical Course of Multiple Sclerosis: The 2013 Revisions." *Neurology* 83(3):278–86.

Lyden, Hannah, Sarah I. Gimbel, Larissa Del Piero, A. Bryna Tsai, Matthew E. Sachs, Jonas T. Kaplan, Gayla Margolin, and Darby Saxbe. 2016. "Associations between Family Adversity and Brain Volume in Adolescence: Manual vs. Automated Brain Segmentation Yields Different Results." *Frontiers in Neuroscience* 10(SEP).

Lysandropoulos, Andreas P., Julie Absil, Thierry Metens, Nicolas Mavroudakis, François Guisset, Eline Van Vlierberghe, Dirk Smeets, Philippe David, Anke Maertens, and Wim Van Hecke. 2016. "Quantifying Brain Volumes for Multiple Sclerosis Patients Follow-up in Clinical Practice - Comparison of 1.5 and 3 Tesla Magnetic Resonance Imaging." *Brain and Behavior* 6(2):1–8.

MAGNIMS. 2021. "MAGNIMS (Magnetic Resonance Imaging in MS)." Retrieved August 23, 2021 (https://www.magnims.eu/).

Manjón, José V. and Pierrick Coupé. 2016. "VolBrain: An Online MRI Brain Volumetry System." *Frontiers in Neuroinformatics* 0(JUL):30.

Martins, Cristina, Nadia Moreira Da Silva, Guilherme Silva, Verena E. Rozanski, and Joao Paulo Silva Cunha. 2016. "Automated Volumetry for Unilateral Hippocampal Sclerosis Detection in Patients with Temporal Lobe Epilepsy." Pp. 6339–42 in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. Vols. 2016-October. Institute of Electrical and Electronics Engineers Inc.

Mathern, G. W., J. K. Pretorius, and T. L. Babb. 1995. "Influence of the Type of Initial Precipitating Injury and at What Age It Occurs on Course and Outcome in Patients with Temporal Lobe Seizures." *Journal of Neurosurgery* 82(2):220–27.

Matsuda, Hiroshi. 2016. "MRI Morphometry in Alzheimer's Disease." *Ageing Research Reviews* 30:17–24.

Maynard, John, Sachi Okuchi, Stephen Wastling, Ayisha Al Busaidi, Ofran Almossawi, Wonderboy Mbatha, Sebastian Brandner, Zane Jaunmuktane, Ali Murat Koc, Laura Mancini, Rolf Jäger, and Stefanie Thust. 2020. "World Health Organization Grade Ii/Iii Glioma Molecular Status: Prediction by Mri Morphologic Features and Apparent Diffusion Coefficient." *Radiology* 296(1):111–21.

McEvoy, Linda K. and James B. Brewer. 2010a. "Quantitative Structural MRI for Early Detection of Alzheimer's Disease." *Expert Review of Neurotherapeutics* 10(11):1675–88.

McEvoy, Linda K. and James B. Brewer. 2010b. "Quantitative Structural MRI for Early Detection of Alzheimer's Disease." *Expert Review of Neurotherapeutics* 10(11):1675.

McGinley, Marisa P., Carolyn H. Goldschmidt, and Alexander D. Rae-Grant. 2021. "Diagnosis and Treatment of Multiple Sclerosis: A Review." *JAMA - Journal of the American Medical Association* 325(8):765–79.

McHugh, Mary L. 2012. "Interrater Reliability: The Kappa Statistic." *Biochemia Medica* 22(3):276–82.

mediaire. 2021. "Mediaire." Retrieved August 23, 2021 (https://mediaire.de/en/home/).

Meier, Raphael, Urspeter Knecht, Tina Loosli, Stefan Bauer, Johannes Slotboom, Roland Wiest, and Mauricio Reyes. 2016. "Clinical Evaluation of a Fully-Automatic Segmentation Method for Longitudinal Brain Tumor Volumetry." 6(1):1–11.

Meiners, Linda C., Ad Van Gils, Gerard H. Jansen, Gerard De Kort, Theo D. Witkamp, Uno M.

P. Ramos, Jaap Valk, Rene M. C. Debets, Alexander C. Van Huffelen, Cees W. M. Van Veelen, and Willem P. T. M. Mali. 1994. “Temporal Lobe Epilepsy: The Various MR Appearances of Histologically Proven Mesial Temporal Sclerosis.” *Am J Neuroradiol* 15:1547–55.

Mendes, Natacha, Sabine Oligschläger, Mark E. Lauckner, Johannes Golchert, Julia M. Huntenburg, Marcel Falkiewicz, Melissa Ellamil, Sarah Krause, Blazej M. Baczkowski, Roberto Cozatl, Anastasia Osoianu, Deniz Kumral, Jared Pool, Laura Golz, Maria Dreyer, Philipp Haueis, Rebecca Jost, Yelyzaveta Kramarenko, Haakon Engen, Katharina Ohrnberger, Krzysztof J. Gorgolewski, Nicolas Farrugia, Anahit Babayan, Andrea Reiter, H. Lina Schaare, Janis Reinelt, Josefin Röbbig, Marie Uhlig, Miray Erbey, Michael Gaebler, Jonathan Smallwood, Arno Villringer, and Daniel S. Margulies. 2019. “A Functional Connectome Phenotyping Dataset Including Cognitive State and Personality Measures.” *Scientific Data 2019 6:1* 6(1):1–19.

Menze, Bjoern H., Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çağatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, José António Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. 2015. “The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS).” *IEEE Transactions on Medical Imaging* 34(10):1993–2024.

Mercado, CL. 2014. “BI-RADS Update.” *Radiologic Clinics of North America* 52(3):481–87.

Mettenburg, J. M., B. F. Branstetter, C. A. Wiley, P. Lee, and R. M. Richardson. 2019. “Improved Detection of Subtle Mesial Temporal Sclerosis: Validation of a Commercially Available Software for Automated Segmentation of Hippocampal Volume.” *American Journal of Neuroradiology* 40(3):440–45.

Miller, David H. and Siobhan M. Leary. 2007. “Primary-Progressive Multiple Sclerosis.” *Lancet Neurology* 6(10):903–12.

Min, Jeeyoung, Won-Jin Moon, Ji Young Jeon, Jin Woo Choi, Yeon-Sil Moon, and Seol-Heui Han. 2017. “Diagnostic Efficacy of Structural MRI in Patients With Mild-to-Moderate Alzheimer Disease: Automated Volumetric Assessment Versus Visual Assessment.” 208(3):617–23.

Mishra, Sundeep. 2017. “FDA, CE Mark or Something Else?-Thinking Fast and Slow.” *Indian Heart Journal* 69(1):1–5.

Mistry, Akshitkumar M., Andrew T. Hale, Lola B. Chambless, Kyle D. Weaver, Reid C. Thompson, and Rebecca A. Ihrie. 2017. “Influence of Glioblastoma Contact with the Lateral Ventricle on Survival: A Meta-Analysis.” *Journal of Neuro-Oncology* 131(1):125–33.

Moggridge, J., S. B. Vos, F. Prados, O. Goodkin, H. G. Pemberton, D. Matthews, P. Schmidt, F. Barkhof, T. A. Yousry, and J. S. Thornton. 2020. “QNICE - Clinical Deployment of Novel Quantitative Image Analysis Techniques Using Python and Containerisation Technology.” in *Artificial Intelligence in MRI Conference. Institute of Physics and Engineering in Medicine*.

Morin, Alexandre, Jorge Samper-Gonzalez, Anne Bertrand, Sébastian Ströer, DIdier Dormont, Aline Mendes, Pierrick Coupé, Jamila Ahdidan, Marcel Lévy, Dalila Samri, Harald Hampel, Bruno Dubois, Marc Teichmann, Stéphane Epelbaum, and Olivier Colliot. 2020. “Accuracy of MRI Classification Algorithms in a Tertiary Memory Center Clinical Routine Cohort.” *Journal of Alzheimer’s Disease* 74(4):1157–66.

NAIMS. 2021. “NAIMS.” Retrieved August 23, 2021 (https://www.naimscooperative.org/).

Namer, I. J., R. Waydelich, J. P. Armspach, E. Hirsch, C. Marescaux, and D. Grucker. 1998. “Contribution of T2 Relaxation Time Mapping in the Evaluation of Cryptogenic Temporal Lobe Epilepsy.” *NeuroImage* 7(4 I):304–13.

NICE. 2018. *Evidence Standards Framework for Digital Health Technologies Corporate Document*.

Nygaard, Gro O., Kristine B. Walhovd, Piotr Sowa, Joy Loi Chepkoech, Atle Bjørnerud, Paulina Due-Tønnessen, Nils I. Landrø, Soheil Damangir, Gabriela Spulber, Andreas B. Storsve, Mona K. Beyer, Anders M. Fjell, Elisabeth G. Celius, and Hanne F. Harbo. 2015. “Cortical Thickness and Surface Area Relate to Specific Symptoms in Early Relapsing-Remitting Multiple Sclerosis.” *Multiple Sclerosis Journal* 21(4):402–14.

O’Rourke, Brian, Wija Oortwijn, and Tara Schuller. 2020. “Announcing the New Definition of Health Technology Assessment.”

Ochs, Alfred L., David E. Ross, Megan D. Zannoni, Tracy J. Abildskov, Erin D. Bigler, and Alzheimer’s Disease Neuroimaging Initiative. 2015. “Comparison of Automated Brain Volume Measures Obtained with NeuroQuant® and FreeSurfer.” *Journal of Neuroimaging* 25(5):721–27.

Okuda, Darin T., Aksel Siva, Orhun Kantarci, Matilde Inglese, Ilana Katz, Melih Tutuncu, B. Mark Keegan, Stacy Donlon, Le H. Hua, Angela Vidal-Jordana, Xavier Montalban, Alex Rovira, Mar Tintoré, Maria Pia Amato, Bruno Brochet, Jérôme De Seze, David Brassat, Patrick Vermersch, Nicola De Stefano, Maria Pia Sormani, Daniel Pelletier, and Christine Lebrun. 2014. “Radiologically Isolated Syndrome: 5-Year Risk for an Initial Clinical Event.” *PLoS ONE* 9(3).

Omoumi, Patrick, Alexis Ducarouge, Antoine Tournier, Hugh Harvey, Charles E. Kahn, Fanny Louvet-de Verchère, Daniel Pinto Dos Santos, Tobias Kober, and Jonas Richiardi. 2021. “To Buy or Not to Buy—Evaluating Commercial AI Solutions in Radiology (the ECLAIR Guidelines).” *European Radiology* 1–11.

Ostrom, Quinn T., Haley Gittleman, Lindsay Stetson, Selene Virk, and Jill S. Barnholtz-Sloan. 2017. “Epidemiology of Intracranial Gliomas.” *Progress in Neurological Surgery* 30:1–11.

Van Paesschen, W., S. Sisodiya, A. Connelly, J. S. Duncan, S. L. Free, A. A. Raymond, R. A. Grünewald, T. Revesz, S. D. Shorvon, D. R. Fish, J. M. Stevens, C. L. Johnson, F. Scaravilli, W. F. J. Harkness, and G. D. Jackson. 1995. “Quantitative Hippocampal MRI and Intractable Temporal Lobe Epilepsy.” *Neurology* 45(12):2233–40.

Van Paesschen, Wim. 2004. “Qualitative and Quantitative Imaging of the Hippocampus in Mesial Temporal Lobe Epilepsy with Hippocampal Sclerosis.” *Neuroimaging Clinics of North America* 14(3):373–400.

Pardini, Matteo, Carole H. Sudre, Ferran Prados, Özgür Yaldizli, Varun Sethi, Nils Muhlert, Rebecca S. Samson, Steven H. Van De Pavert, M. Jorge Cardoso, Sebastien Ourselin, Claudia A. M. Gandini Wheeler-Kingshott, David H. Miller, and Declan T. Chard. 2016. “Relationship of Grey and White Matter Abnormalities with Distance from the Surface of the Brain in Multiple Sclerosis.” *Journal of Neurology, Neurosurgery and Psychiatry* 87(11):1212–17.

Pareto, Deborah, Jaume Sastre-Garriga, Manel Alberich, Cristina Auger, Mar Tintoré, Xavier Montalban, and Àlex Rovira. 2019. “Brain Regional Volume Estimations with NeuroQuant and FIRST: A Study in Patients with a Clinically Isolated Syndrome.” *Neuroradiology* 61(6):667–74.

Park, Chae Jung, Kyunghwa Han, Haesol Shin, Sung Soo Ahn, Yoon Seong Choi, Yae Won Park, Jong Hee Chang, Se Hoon Kim, Rajan Jain, and Seung-Koo Koo Lee. 2020. “MR Image Phenotypes May Add Prognostic Value to Clinical Features in IDH Wild-Type Lower-Grade Gliomas.” 30(6):3035–45.

Park, Y. W. YW, K. Han, S. S. S. Ahn, S. Bae, Y. S. S. Choi, J. H. H. Chang, S. H. H. Kim, S. G. G. Kang, and S. K. K. Lee. 2018. “Prediction of IDH1 -Mutation and 1p/19q-Codeletion Status Using Preoperative MR Imaging Phenotypes in Lower Grade Gliomas.” *American Journal of Neuroradiology* 39(1):37–42.

Patil, Chirag G., Anthony Yi, Adam Elramsisy, Jethro Hu, Debraj Mukherjee, Dwain K. Irvin, John S. Yu, Serguei I. Bannykh, Keith L. Black, and Miriam Nuño. 2012. “Prognosis of Patients with Multifocal Glioblastoma: A Case-Control Study.” *Journal of Neurosurgery* 117(4):705–11.

Peixoto-Santos, Jose Eduardo, Ludmyla Kandratavicius, Tonicarlo Rodrigues Velasco, Joao Alberto Assirati, Carlos Gilberto Carlotti, Renata Caldo Scandiuzzi, Carlos Ernesto Garrido Salmon, Antonio Carlos dos Santos, and Joao Pereira Leite. 2017. “Individual Hippocampal Subfield Assessment Indicates That Matrix Macromolecules and Gliosis Are Key Elements for the Increased T2 Relaxation Time Seen in Temporal Lobe Epilepsy.” *Epilepsia* 58(1):149–59.

Pereira, J. B., L. Cavallin, G. Spulber, C. Aguilar, P. Mecocci, B. Vellas, M. Tsolaki, I. Kłoszewska, H. Soininen, C. Spenger, D. Aarsland, S. Lovestone, A. Simmons, L. O. Wahlund, E. Westman, and AddNeuroMed consortium and for the Alzheimer’s Disease Neuroimaging Initiative. 2014. “Influence of Age, Disease Onset and *ApoE4* on Visual Medial Temporal Lobe Atrophy Cut-Offs.” *Journal of Internal Medicine* 275(3):317–30.

Pereira, Sergio, Adriano Pinto, Victor Alves, and Carlos A. Silva. 2016. “Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images.” *IEEE Transactions on Medical Imaging* 35(5):1240–51.

Pérez-Miralles, F., J. Sastre-Garriga, M. Tintoré, G. Arrambide, C. Nos, H. Perkal, J. Río, M. C. Edo, A. Horga, J. Castilló, C. Auger, E. Huerga, A. Rovira, and X. Montalban. 2013. “Clinical Impact of Early Brain Atrophy in Clinically Isolated Syndromes.” *Multiple Sclerosis Journal* 19(14):1878–86.

Persson, Karin, Maria Lage Barca, Lena Cavallin, Anne Brækhus, Anne Brita Knapskog, Geir Selbæk, and Knut Engedal. 2018. “Comparison of Automated Volumetry of the

Hippocampus Using NeuroQuant® and Visual Assessment of the Medial Temporal Lobe in Alzheimer's Disease." *Acta Radiologica* 59(8):997–1001.

Persson, Karin, Geir Selbæk, Anne Brækhus, Mona Beyer, Maria Barca, and Knut Engedal. 2017. "Fully Automated Structural MRI of the Brain in Clinical Dementia Workup." *Acta Radiologica* 58(6):740–47.

Pichler, A., M. Khalil, C. Langkammer, D. Pinter, G. Bachmaier, S. Ropele, S. Fuchs, C. Enzinger, and F. Fazekas. 2016. "Combined Analysis of Global and Compartmental Brain Volume Changes in Early Multiple Sclerosis in Clinical Practice." 22(3):340–46.

Polman, Chris H., Stephen C. Reingold, Brenda Banwell, Michel Clanet, Jeffrey A. Cohen, Massimo Filippi, Kazuo Fujihara, Eva Havrdova, Michael Hutchinson, Ludwig Kappos, Fred D. Lublin, Xavier Montalban, Paul O'Connor, Magnhild Sandberg-Wollheim, Alan J. Thompson, Emmanuelle Waubant, Brian Weinshenker, and Jerry S. Wolinsky. 2011. "Diagnostic Criteria for Multiple Sclerosis: 2010 Revisions to the McDonald Criteria." *Annals of Neurology* 69(2):292–302.

Popescu, Veronica, Roel Klaver, Adriaan Versteeg, Pieter Voorn, Jos W. R. Twisk, Frederik Barkhof, Jeroen J. G. Geurts, and Hugo Vrenken. 2016. "Postmortem Validation of MRI Cortical Volume Measurements in MS." *Human Brain Mapping* 37(6):2223–33.

Porz, Nicole, Stefan Bauer, Alessia Pica, Philippe Schucht, Jürgen Beck, Rajeev Kumar Verma, Johannes Slotboom, Mauricio Reyes, and Roland Wiest. 2014. "Multi-Modal Glioblastoma Segmentation: Man versus Machine." *PLOS ONE* 9(5):e96873.

Prados, Ferran, Manuel Jorge Cardoso, Baris Kanber, Olga Ciccarelli, Raju Kapoor, Claudia A. M. Gandini Wheeler-Kingshott, and Sebastien Ourselin. 2016. "A Multi-Time-Point Modality-Agnostic Patch-Based Method for Lesion Filling in Multiple Sclerosis." *NeuroImage* 139:376–84.

Prados, Ferran, Manuel Jorge Cardoso, Kelvin K. Leung, David M. Cash, Marc Modat, Nick C. Fox, Claudia A. M. Wheeler-Kingshott, and Sebastien Ourselin. 2015. "Measuring Brain Atrophy with a Generalized Formulation of the Boundary Shift Integral." *Neurobiology of Aging* 36(S1):S81–90.

Prince, M., M. Knapp, M. Guerchet, P. McCrone, M. Prina, A. Comas-Herrera, R. Wittenberg, B. Adelaja, B. Hu, D. King, A. Rehill, and D. Salimkumar. 2014. "Dementia UK: Update Second Edition." Retrieved May 21, 2020 (https://www.alzheimers.org.uk/sites/default/files/migrate/downloads/dementia_uk_update.pdf).

Quantib. 2021. "Artificial Intelligence in Healthcare & Radiology | Quantib." Retrieved August 23, 2021 (https://www.quantib.com/).

Raji, Cyrus A., Maria Ly, and Tammie L. S. Benzinger. 2019. "Overview of MR Imaging Volumetric Quantification in Neurocognitive Disorders." *Topics in Magnetic Resonance Imaging* 28(6):311–15.

Recht, Michael P., Marc Dewey, Keith Dreyer, Curtis Langlotz, Wiro Niessen, Barbara Prainsack, and John J. Smith. 2020. "Integrating Artificial Intelligence into the Clinical Practice of Radiology: Challenges and Recommendations." *European Radiology* 30(6):3576–84.

Reutens, D. C., J. M. Stevens, D. Kingsley, B. Kendall, I. Moseley, M. J. Cook, S. Free, D. R. Fish, and S. D. Shorvon. 1996. "Reliability of Visual Inspection for Detection of

Volumetric Hippocampal Asymmetry." *Neuroradiology* 38(3):221–25.

Rezazade Mehrizi, MH, P. van Ooijen, and M. Homan. 2021. "Applications of Artificial Intelligence (AI) in Diagnostic Radiology: A Technography Study." *European Radiology* 31(4):1805–11.

Risacher, Shannon L. and Andrew J. Saykin. 2013. "Neuroimaging Biomarkers of Neurodegenerative Diseases and Dementia." *Seminars in Neurology* 33(4):386–416.

Risacher, Shannon L., Andrew J. Saykin, John D. West, Li Shen, Hiram A. Firpi, Brenna C. McDonald, and Alzheimer's Disease Neuroimaging Initiative (ADNI). 2009. "Baseline MRI Predictors of Conversion from MCI to Probable AD in the ADNI Cohort." *Current Alzheimer Research* 6(4):347–61.

Rocca, Maria A., Marco Battaglini, Ralph H. B. Benedict, Nicola De Stefano, Jeroen J. G. Geurts, Roland G. Henry, Mark A. Horsfield, Mark Jenkinson, Elisabetta Pagani, and Massimo Filippi. 2017. "Brain MRI Atrophy Quantification in MS." *Neurology* 88(4):403–13.

Rodionov, R., P. A. A. Bartlett, Ci He, S. B. B. Vos, N. K. K. Focke, S. G. G. Ourselin, and J. S. S. Duncan. 2015. "T2 Mapping Outperforms Normalised FLAIR in Identifying Hippocampal Sclerosis." *NeuroImage: Clinical* 7:788–91.

Ross, David E., Alfred L. Ochs, Megan E. Desmit, Jan M. Seabaugh, and Micha El D. Havranek. 2015. "Man versus Machine Part 2: Comparison of Radiologists' Interpretations and Neuroquant Measures of Brain Asymmetry and Progressive Atrophy in Patients with Traumatic Brain Injury." *Journal of Neuropsychiatry and Clinical Neurosciences* 27(2):147–52.

Ross, David E., Alfred L. Ochs, Jan M. Seabaugh, Carole R. Shrader, and Alzheimer's Disease Neuroimaging Initiative. 2013. "Man Versus Machine: Comparison of Radiologists' Interpretations and NeuroQuant ® Volumetric Analyses of Brain MRIs in Patients With Traumatic Brain Injury." *The Journal of Neuropsychiatry and Clinical Neurosciences* 25(1):32–39.

Rossor, Martin N., Nick C. Fox, Catherine J. Mummery, Jonathan M. Schott, and Jason D. Warren. 2010. "The Diagnosis of Young-Onset Dementia." *The Lancet Neurology* 9(8):793–806.

Rovaris, Marco, Christian Confavreux, Roberto Furlan, Ludwig Kappos, Giancarlo Comi, and Massimo Filippi. 2006. "Secondary Progressive Multiple Sclerosis: Current Knowledge and Future Challenges." *Lancet Neurology* 5(4):343–54.

RSNA. 2007. "Quantitative Imaging Biomarkers Alliance." Retrieved August 23, 2021 (https://www.rsna.org/research/quantitative-imaging-biomarkers-alliance).

Rubin, DL. 2019. "Artificial Intelligence in Imaging: The Radiologist's Role." *Journal of the American College of Radiology : JACR* 16(9 Pt B):1309–17.

Salvatore, Christian, Antonio Cerasa, and Isabella Castiglioni. 2018. "MRI Characterizes the Progressive Course of AD and Predicts Conversion to Alzheimer's Dementia 24 Months Before Probable Diagnosis." *Frontiers in Aging Neuroscience* 10:135.

Sand, Ilana Katz. 2015. "Classification, Diagnosis, and Differential Diagnosis of Multiple Sclerosis." *Current Opinion in Neurology* 28(3):193–205.

Saslow, Lori, David K. B. Li, June Halper, Brenda Banwell, Frederik Barkhof, Laura Barlow,

Kathleen Costello, Peter Damiri, Jeffrey Dunn, Shivraman Giri, Micki Maes, Sarah Morrow, Scott Newsome, Jiwon Oh, Friedemann Paul, Patrick Quarterman, Daniel Reich, Jason R. Shewchuk, Russel Takeshi Shinohara, Wim Van Hecke, Kim van de Ven, Mitchell Wallin, Jerry Wolinsky, and Anthony Traboulsee. 2020. "An International Standardized Magnetic Resonance Imaging Protocol for Diagnosis and Follow-up of Patients with Multiple Sclerosis: Advocacy, Dissemination and Implementation Strategies." *International Journal of MS Care* In press.

Sastre-Garriga, Jaume, Deborah Pareto, Marco Battaglini, Maria A. Rocca, Olga Ciccarelli, Christian Enzinger, Jens Wuerfel, Maria P. Sormani, Frederik Barkhof, Tarek A. Yousry, Nicola De Stefano, Mar Tintoré, Massimo Filippi, Claudio Gasperini, Ludwig Kappos, Jordi Río, Jette Frederiksen, Jackie Palace, Hugo Vrenken, Xavier Montalban, and Àlex Rovira. 2020. "MAGNIMS Consensus Recommendations on the Use of Brain and Spinal Cord Atrophy Measures in Clinical Practice." *Nature Reviews Neurology* 16(3):171–82.

Sastre-Garriga, Jaume, Deborah Pareto, and Àlex Rovira. 2017. "Brain Atrophy in Multiple Sclerosis: Clinical Relevance and Technical Aspects." *Neuroimaging Clinics of North America* 27(2):289–300.

Scalfari, Antonio, Chiara Romualdi, Richard S. Nicholas, Miriam Mattoscio, Roberta Magliozzi, Aldo Morra, Salvatore Monaco, Paolo A. Muraro, and Massimiliano Calabrese. 2018. "The Cortical Damage, Early Relapses, and Onset of the Progressive Phase in Multiple Sclerosis." *Neurology* 90(24):e2099–2106.

Scheek, Damian, Mohammad. H. Rezazade Mehrizi, and Erik Ranschaert. 2021. "Radiologists in the Loop: The Roles of Radiologists in the Development of AI Applications." *European Radiology* 1–9.

Scheltens, P., D. Leys, F. Barkhof, D. Huglo, H. C. Weinstein, P. Vermersch, M. Kuiper, M. Steinling, E. C. Wolters, and J. Valk. 1992. "Atrophy of Medial Temporal Lobes on MRI in "Probable" Alzheimer's Disease and Normal Ageing: Diagnostic Value and Neuropsychological Correlates." *Journal of Neurology, Neurosurgery, and Psychiatry* 55(10):967–72.

Schmierer, Klaus, Thomas Campion, Audrey Sinclair, Wim Van Hecke, Paul M. Matthews, and Mike P. Wattjes. 2019. "Commentary: Towards a Standard MRI Protocol for Multiple Sclerosis across the UK." *British Journal of Radiology* 92(1101).

Schwartz, Lawrence H., David M. Panicek, Alexandra R. Berk, Yuelin Li, and Hedvig Hricak. 2011. "Improving Communication of Diagnostic Radiology Findings through Structured Reporting." *Radiology* 260(1):174–81.

Schwarz, Adam J., Karen L. Sundell, Arnaud Charil, Michael G. Case, Ralf K. Jaeger, D. Scott, Luc Bracoud, Joonmi Oh, J. Suhy, Michael J. Pontecorvo, Bradford C. Dickerson, and Eric R. Siemers. 2019. "Magnetic Resonance Imaging Measures of Brain Atrophy from the EXPEDITION3 Trial in Mild Alzheimer's Disease." *Alzheimer's and Dementia: Translational Research and Clinical Interventions* 5(1):328–37.

Shen, Guiquan, Rujia Wang, Bo Gao, Zhongwen Zhang, Guipeng Wu, and Whitney Pope. 2020. "The MRI Features and Prognosis of Gliomas Associated With IDH1 Mutation: A Single Center Study in Southwest China." *Frontiers in Oncology* 0:852.

Sidhu, Meneka Kaur, John S. Duncan, and Josemir W. Sander. 2018. "Neuroimaging in Epilepsy." *Current Opinion in Neurology* 31(4):371–78.

Simões, Rita, Christoph Mönninghoff, Martha Dlugaj, Christian Weimar, Isabel Wanke, Anne

Marie van Cappellen van Walsum, and Cornelis Slump. 2013. “Automatic Segmentation of Cerebral White Matter Hyperintensities Using Only 3D FLAIR Images.” *Magnetic Resonance Imaging* 31(7):1182–89.

Sled, JG, AP Zijdenbos, and AC Evans. 1998. “A Nonparametric Method for Automatic Correction of Intensity Nonuniformity in MRI Data.” *IEEE Transactions on Medical Imaging* 17(1):87–97.

Smith, E. T. S. 2011. “Clinical Applications of Imaging Biomarkers. Part 1. The Neuroradiologist’s Perspective.” *The British Journal of Radiology* 84(Spec Iss 2):S196.

Smith, Stephen M., Yongyue Zhang, Mark Jenkinson, Jacqueline Chen, P. M. Matthews, Antonio Federico, and Nicola De Stefano. 2002. “Accurate, Robust, and Automated Longitudinal and Cross-Sectional Brain Change Analysis.” *NeuroImage* 17(1):479–89.

Smits, Marion. 2021. “MRI Biomarkers in Neuro-Oncology.” *Nature Reviews Neurology*.

Solomon, Alina, Miia Kivipelto, José Luis Molinuevo, Brian Tom, and Craig W. Ritchie. 2018. “European Prevention of Alzheimer’s Dementia Longitudinal Cohort Study (EPAD LCS): Study Protocol.” *BMJ Open* 8(12):e021017.

Sormani, Maria Pia, Ludwig Kappos, Ernst Wilhelm Radue, Jeffrey Cohen, Frederik Barkhof, Till Sprenger, Daniela Piani Meier, Dieter Häring, Davorka Tomic, and Nicola De Stefano. 2017. “Defining Brain Volume Cutoffs to Identify Clinically Relevant Atrophy in RRMS.” *Multiple Sclerosis* 23(5):656–64.

Staffaroni, Adam M., Fanny M. Elahi, Dana McDermott, Kacey Marton, Elissaios Karageorgiou, Simone Sacco, Matteo Paoletti, Eduardo Caverzasi, Christopher P. Hess, Howard J. Rosen, and Michael D. Geschwind. 2017. “Neuroimaging in Dementia.” *Seminars in Neurology* 37(5):510–37.

Steenwijk, Martijn D., Houshang Amiri, Menno M. Schoonheim, Alexandra de Sitter, Frederik Barkhof, Petra J. W. Pouwels, and Hugo Vrenken. 2017. “Agreement of MSmetrix with Established Methods for Measuring Cross-Sectional and Longitudinal Brain Atrophy.” *NeuroImage: Clinical* 15:843–53.

De Stefano, N., P. M. Matthews, M. Filippi, F. Agosta, M. De Luca, M. L. Bartolozzi, L. Guidi, A. Ghezzi, E. Montanari, A. Cifelli, A. Federico, and S. M. Smith. 2003. “Evidence of Early Cortical Atrophy in MS: Relevance to White Matter Changes and Disability.” *Neurology* 60(7):1157–62.

De Stefano, Nicola, Antonio Giorgio, Mar Tintoré, Maria Pia Amato, Ludwig Kappos, Jacqueline Palace, Tarek Yousry, Maria A. Rocca, Olga Ciccarelli, Christian Enzinger, Jette Frederiksen, Massimo Filippi, Hugo Vrenken, and Àlex Rovira. 2018. “Radiologically Isolated Syndrome or Subclinical Multiple Sclerosis: MAGNIMS Consensus Recommendations.” *Multiple Sclerosis* 24(2):214–21.

De Stefano, Nicola, Maria Laura Stromillo, Antonio Giorgio, Maria Letizia Bartolozzi, Marco Battaglini, Mariella Baldini, Emilio Portaccio, Maria Pia Amato, and Maria Pia Sormani. 2016. “Establishing Pathological Cut-Offs of Brain Atrophy Rates in Multiple Sclerosis.” *Journal of Neurology, Neurosurgery, and Psychiatry* 87(1):93–99.

Strohm, L., Hehakaya C, Ranschaert ER, Boon WPC, and Moors EHM. 2020. “Implementation of Artificial Intelligence (AI) Applications in Radiology: Hindering and Facilitating Factors.” *European Radiology* 30(10):5525–32.

Struyfs, Hanne, Diana Maria Sima, Melissa Wittens, Annemie Ribbens, Nuno Pedrosa de

Barros, Thanh Vân Phan, Maria Ines Ferraz Meyer, Lene Claes, Ellis Niemantsverdriet, Sebastiaan Engelborghs, Wim Van Hecke, and Dirk Smeets. 2020. “Automated MRI Volumetry as a Diagnostic Tool for Alzheimer’s Disease: Validation of Icobrain Dm.” *NeuroImage: Clinical* 26:102243.

Stupp, Roger, Warren P. Mason, Martin J. van den Bent, Michael Weller, Barbara Fisher, Martin J. B. Taphoorn, Karl Belanger, Alba A. Brandes, Christine Marosi, Ulrich Bogdahn, Jürgen Curschmann, Robert C. Janzer, Samuel K. Ludwin, Thierry Gorlia, Anouk Allgeier, Denis Lacombe, J. Gregory Cairncross, Elizabeth Eisenhauer, and René O. Mirimanoff. 2005. “Radiotherapy plus Concomitant and Adjuvant Temozolomide for Glioblastoma.” *New England Journal of Medicine* 352(10):987–96.

Su, C. Q., S. S. Lu, M. D. Zhou, H. Shen, H. B. Shi, and X. N. Hong. 2019. “Combined Texture Analysis of Diffusion-Weighted Imaging with Conventional MRI for Non-Invasive Assessment of IDH1 Mutation in Anaplastic Gliomas.” *Clinical Radiology* 74(2):154–60.

Sudre, C. H., B. Gomez Anson, I. Davagnanam, A. Schmitt, A. F. Mendelson, F. Prados, L. Smith, D. Atkinson, A. D. Hughes, N. Chaturvedi, M. J. Cardoso, F. Barkhof, H. R. Jaeger, and S. Ourselin. 2018. “Bullseye’s Representation of Cerebral White Matter Hyperintensities.” *Journal of Neuroradiology. Journal de Neuroradiologie* 45(2):114–22.

Sudre, Carole H., M. Jorge Cardoso, Willem H. Bouvy, Geert Jan Biessels, Josephine Barnes, and Sebastien Ourselin. 2015. “Bayesian Model Selection for Pathological Neuroimaging Data Applied to White Matter Lesion Segmentation.” *IEEE Transactions on Medical Imaging* 34(10):2079–2102.

Sullivan, Daniel C., Nancy A. Obuchowski, Larry G. Kessler, David L. Raunig, Constantine Gatsonis, Erich P. Huang, Marina Kondratovich, Lisa M. McShane, Anthony P. Reeves, Daniel P. Barboriak, Alexander R. Guimaraes, Richard L. Wahl, and RSNA-QIBA Metrology Working Group. 2015. “Metrology Standards for Quantitative Imaging Biomarkers.” *Radiology* 277(3):813–25.

Suzuki, Hiromichi, Kosuke Aoki, Kenichi Chiba, Yusuke Sato, Yusuke Shiozawa, Yuichi Shiraishi, Teppei Shimamura, Atsushi Niida, Kazuya Motomura, Fumiharu Ohka, Takashi Yamamoto, Kuniaki Tanahashi, Melissa Ranjit, Toshihiko Wakabayashi, Tetsuichi Yoshizato, Keisuke Kataoka, Kenichi Yoshida, Yasunobu Nagata, Aiko Sato-Otsubo, Hiroko Tanaka, Masashi Sanada, Yutaka Kondo, Hideo Nakamura, Masahiro Mizoguchi, Tatsuya Abe, Yoshihiro Muragaki, Reiko Watanabe, Ichiro Ito, Satoru Miyano, Atsushi Natsume, and Seishi Ogawa. 2015. “Mutational Landscape and Clonal Architecture in Grade II and III Gliomas.” *Nature Genetics* 47(5):458–68.

Tanpitukpongse, T. P., M. A. Mazurowski, J. Ikhena, and J. R. Petrella. 2017. “Predictive Utility of Marketed Volumetric Software Tools in Subjects at Risk for Alzheimer Disease: Do Regions Outside the Hippocampus Matter?” *American Journal of Neuroradiology* 38(3):546–52.

Thom, Maria, Sofia Eriksson, Lillian Martinian, Luis O. Caboclo, Andrew W. McEvoy, John S. Duncan, and Sanjay M. Sisodiya. 2009. “Temporal Lobe Sclerosis Associated With Hippocampal Sclerosis in Temporal Lobe Epilepsy: Neuropathological Features.” *Journal of Neuropathology & Experimental Neurology* 68(8):928–38.

Thompson, Alan J., Sergio E. Baranzini, Jeroen Geurts, Bernhard Hemmer, and Olga Ciccarelli. 2018. “Multiple Sclerosis.” *The Lancet* 391(10130):1622–36.

Thust, S. C., S. Heiland, A. Falini, H. R. Jäger, A. D. Waldman, P. C. Sundgren, C. Godi, V. K.

Katsaros, A. Ramos, N. Bargallo, M. W. Vernooij, T. Yousry, M. Bendszus, and M. Smits. 2018. “Glioma Imaging in Europe: A Survey of 220 Centres and Recommendations for Best Clinical Practice.” *European Radiology* 28(8):3306.

Tillin, Therese, Nita G. Forouhi, Paul M. McKeigue, Nish Chaturvedi, Nish Chaturvedi, Norman Beauchamp, Emma Coady, Rory Collins, Nita Forouhi, Wladyslaw Gedroyc, Ian Godsland, Andrew Hattersley, Alun Hughes, Farrukh Majeed, Jamil Mayet, Paul McKeigue, Naveed Sattar, Dean Shibata, Peter Whincup, and Andrew Wright. 2012. “Southall And Brent REvisited: Cohort Profile of SABRE, a UK Population-Based Comparison of Cardiovascular Disease and Diabetes in People of European, Indian Asian and African Caribbean Origins.” *International Journal of Epidemiology* 41(1):33–42.

UCSF. 2021. “Research Studies at the MAC | Memory and Aging Center.” Retrieved August 23, 2021 (https://memory.ucsf.edu/research-trials/research#Studies-for-People-with-Frontotemporal-Spectrum-Disorders-including-FTD-PPA-CBS-PSP).

Valverde, S. 2021. “GitHub - Sergivalverde/NicMSlesions: Easy Multiple Sclerosis White Matter Lesion Segmentation Using Convolutional Deep Neural Networks.” Retrieved August 23, 2021 (https://github.com/sergivalverde/nicMSlesions/).

Valverde, S., A. Oliver, Y. Díez, M. Cabezas, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira, and X. Lladó. 2015. “Evaluating the Effects of White Matter Multiple Sclerosis Lesions on the Volume Estimation of 6 Brain Tissue Segmentation Methods.” *American Journal of Neuroradiology* 36(6):1109–15.

Valverde, Sergi, Mariano Cabezas, Eloy Roura, Sandra González-Villà, Deborah Pareto, Joan C. Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Arnau Oliver, and Xavier Lladó. 2017. “Improving Automated Multiple Sclerosis Lesion Segmentation with a Cascaded 3D Convolutional Neural Network Approach.” *NeuroImage* 155:159–68.

Vandenberghe, Rik, Marie Emmanuelle Riviere, Angelika Caputo, Judit Sovago, R. Paul Maguire, Martin Farlow, Giovanni Marotta, Raquel Sanchez-Valle, Philip Scheltens, J. Michael Ryan, and Ana Graf. 2017. “Active Aβ Immunotherapy CAD106 in Alzheimer’s Disease: A Phase 2b Study.” *Alzheimer’s and Dementia: Translational Research and Clinical Interventions* 3(1):10–22.

Vernooij, M. W., F. B. Pizzini, R. Schmidt, M. Smits, T. A. Yousry, N. Bargallo, G. B. Frisoni, S. Haller, F. Barkhof, and M. W. Vernooij mvernooij. 2019. “Dementia Imaging in Clinical Practice: A European-Wide Survey of 193 Centres and Conclusions by the ESNR Working Group.” *Neuroradiology* 61(6):633–42.

Vernooij, Meike W., Bas Jasperse, Rebecca Steketee, Marcel Koek, Henri Vrooman, M. Arfan Ikram, Janne Papma, Aad van der Lugt, Marion Smits, and Wiro J. Niessen. 2018. “Automatic Normative Quantification of Brain Tissue Volume to Support the Diagnosis of Dementia: A Clinical Evaluation of Diagnostic Accuracy.” *NeuroImage: Clinical* 20(July):374–79.

Vernooij, Meike W. and Marion Smits. 2012. *Structural Neuroimaging in Aging and Alzheimer’s Disease*. Vol. 22. Neuroimaging Clin N Am.

Vinke, Elisabeth J., Wyke Huizinga, Martin Bergtholdt, Hieab H. Adams, Rebecca M. E. Steketee, Janne M. Papma, Frank J. de Jong, Wiro J. Niessen, M. Arfan Ikram, Fabian Wenzel, and Meike W. Vernooij. 2019. “Normative Brain Volumetry Derived from Different Reference Populations: Impact on Single-Subject Diagnostic Assessment in Dementia.” *Neurobiology of Aging* 84:9–16.

Vos, M. J., B. M. J. Uitdehaag, F. Barkhof, J. J. Heimans, H. C. Baayen, W. Boogerd, J. A. Castelijns, P. H. M. Elkhuizen, and T. J. Postma. 2003. “Interobserver Variability in the Radiological Assessment of Response to Chemotherapy in Glioma.” *Neurology* 60(5):826–30.

Vos, Sjoerd B., Caroline Micallef, Frederik Barkhof, Andrea Hill, Gavin P. Winston, Sebastien Ourselin, and John S. Duncan. 2018. “Evaluation of Prospective Motion Correction of High-Resolution 3D-T2-FLAIR Acquisitions in Epilepsy Patients.” *Journal of Neuroradiology* 45(6):368–73.

Vos, Sjoerd B., Gavin P. Winston, Olivia Goodkin, Hugh G. Pemberton, Frederik Barkhof, Ferran Prados, Marian Galovic, Matthias Koepp, Sebastien Ourselin, M. Jorge Cardoso, and John S. Duncan. 2019. “Hippocampal Profiling: Localized Magnetic Resonance Imaging Volumetry and T2 Relaxometry for Hippocampal Sclerosis.” *Epilepsia*.

Wang, Huiquan, S. Nizam Ahmed, and Mrinal Mandal. 2020. “Automated Detection of Focal Cortical Dysplasia Using a Deep Convolutional Neural Network.” *Computerized Medical Imaging and Graphics* 79:101662.

Wangaryattawanich, Pattana, Masumeh Hatami, Jixin Wang, Ginu Thomas, Adam Flanders, Justin Kirby, Max Wintermark, Erich S. Huang, Ali Shojaee Bakhtiari, Markus M. Luedi, Syed S. Hashmi, Daniel L. Rubin, James Y. Chen, Scott N. Hwang, John Freymann, Chad A. Holder, Pascal O. Zinn, and Rivka R. Colen. 2015. “Multicenter Imaging Outcomes Study of The Cancer Genome Atlas Glioblastoma Patient Cohort: Imaging Predictors of Overall and Progression-Free Survival.” *Neuro-Oncology* 17(11):1525–37.

Warfield, Simon K., Kelly H. Zou, and William M. Wells. 2004. “Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation.” *IEEE Transactions on Medical Imaging* 23(7):903–21.

Wattjes, Mike P. 2011. “Structural MRI.” *International Psychogeriatrics* 23(S2):S13–24.

Wattjes, Mike P., Olga Ciccarelli, Daniel S. Reich, Brenda Banwell, Nicola de Stefano, Christian Enzinger, Franz Fazekas, Massimo Filippi, Jette Frederiksen, Claudio Gasperini, Yael Hacohen, Ludwig Kappos, David K. B. Li, Kshitij Mankad, Xavier Montalban, Scott D. Newsome, Jiwon Oh, Jacqueline Palace, Maria A Rocca, Jaume Sastre-Garriga, Mar Tintoré, Anthony Traboulsee, Hugo Vrenken, Tarek Yousry, Frederik Barkhof, Àlex Rovira, Mike P. Wattjes, Olga Ciccarelli, Nicola de Stefano, Christian Enzinger, Franz Fazekas, Massimo Filippi, Jette Frederiksen, Claudio Gasperini, Yael Hacohen, Ludwig Kappos, Kshitij Mankad, Xavier Montalban, Jacqueline Palace, María A Rocca, Jaume Sastre-Garriga, Mar Tintore, Hugo Vrenken, Tarek Yousry, Frederik Barkhof, Alex Rovira, David K. B. Li, Anthony Traboulsee, Scott D. Newsome, Brenda Banwell, Jiwon Oh, Daniel S. Reich, Daniel S. Reich, and Jiwon Oh. 2021. “2021 MAGNIMS–CMSC–NAIMS Consensus Recommendations on the Use of MRI in Patients with Multiple Sclerosis.” *The Lancet Neurology* 0(0).

Wattjes, Mike P., Àlex Rovira, David Miller, Tarek A. Yousry, Maria P. Sormani, Nicola De Stefano, Mar Tintoré, Cristina Auger, Carmen Tur, Massimo Filippi, Maria A. Rocca, Franz Fazekas, Ludwig Kappos, Chris Polman, Frederik Barkhof, and Xavier Montalban. 2015. “MAGNIMS Consensus Guidelines on the Use of MRI in Multiple Sclerosis—Establishing Disease Prognosis and Monitoring Patients.” *Nature Reviews Neurology* 11(10):597–606.

Weiss, D. L. and C. P. Langlotz. 2008. “Structured Reporting: Patient Care Enhancement or Productivity Nightmare?” *Radiology* 249:739–47.

Wellmer, Jörg, Carlos M. Quesada, Lars Rothe, Christian E. Elger, Christian G. Bien, and Horst Urbach. 2013. “Proposal for a Magnetic Resonance Imaging Protocol for the Detection of Epileptogenic Lesions at Early Outpatient Stages.” *Epilepsia* 54(11):1977–87.

Wen, Patrick Y., David R. Macdonald, David A. Reardon, Timothy F. Cloughesy, A. Gregory Sorensen, Evanthia Galanis, John DeGroot, Wolfgang Wick, Mark R. Gilbert, Andrew B. Lassman, Christina Tsien, Tom Mikkelsen, Eric T. Wong, Marc C. Chamberlain, Roger Stupp, Kathleen R. Lamborn, Michael A. Vogelbaum, Martin J. van den Bent, and Susan M. Chang. 2010. “Updated Response Assessment Criteria for High-Grade Gliomas: Response Assessment in Neuro-Oncology Working Group.” *Journal of Clinical Oncology* 28(11):1963–72.

Whitwell, J. L., M. M. Shiung, S. A. Przybelski, S. D. Weigand, D. S. Knopman, B. F. Boeve, R. C. Petersen, and C. R. Jack. 2008. “MRI Patterns of Atrophy Associated with Progression to AD in Amnestic Mild Cognitive Impairment.” *Neurology* 70(7):512–20.

Wijnenga, Maarten M. J., Sebastian R. van der Voort, Pim J. French, Stefan Klein, Hendrikus J. Dubbink, Winand N. M. Dinjens, Peggy N. Atmodimedjo, Marius de Groot, Johan M. Kros, Joost W. Schouten, Clemens M. F. Dirven, Arnaud J. P. E. Vincent, Marion Smits, and Martin J. van den Bent. 2019. “Differences in Spatial Distribution between WHO 2016 Low-Grade Glioma Molecular Subgroups.” *Neuro-Oncology Advances* 1(1):1–9.

wiki.cancerimagingarchive.net. n.d. “VASARI Research Project - The Cancer Imaging Archive (TCIA) Public Access - Cancer Imaging Archive Wiki.” Retrieved July 31, 2020 (https://wiki.cancerimagingarchive.net/display/Public/VASARI+Research+Project).

Wilkinson, Beata and Robert van Boxtel. 2019. “The Medical Device Regulation of the European Union Intensifies Focus on Clinical Benefits of Devices.” *Therapeutic Innovation & Regulatory Science* 216847901987073.

Winblad, Bengt, Philippe Amouyel, Sandrine Andrieu, Clive Ballard, Carol Brayne, Henry Brodaty, Angel Cedazo-Minguez, Bruno Dubois, David Edvardsson, Howard Feldman, Laura Fratiglioni, Giovanni B. Frisoni, Serge Gauthier, Jean Georges, Caroline Graff, Khalid Iqbal, Frank Jessen, Gunilla Johansson, Linus Jönsson, Miia Kivipelto, Martin Knapp, Francesca Mangialasche, René Melis, Agneta Nordberg, Marcel Olde Rikkert, Chengxuan Qiu, Thomas P. Sakmar, Philip Scheltens, Lon S. Schneider, Reisa Sperling, Lars O. Tjernberg, Gunhild Waldemar, Anders Wimo, and Henrik Zetterberg. 2016. “Defeating Alzheimer’s Disease and Other Dementias: A Priority for European Science and Society.” *The Lancet. Neurology* 15(5):455–532.

Winston, Gavin P., M. Jorge Cardoso, Elaine J. Williams, Jane L. Burdett, Philippa A. Bartlett, Miklos Espak, Charles Behr, John S. Duncan, and Sebastien Ourselin. 2013. “Automated Hippocampal Segmentation in Patients with Epilepsy: Available Free Online.” *Epilepsia* 54(12):2166–73.

Winston, Gavin P., Sjoerd B. Vos, Jane L. Burdett, M. Jorge Cardoso, Sebastien Ourselin, and John S. Duncan. 2017. “Automated T2 Relaxometry of the Hippocampus for Temporal Lobe Epilepsy.” *Epilepsia* 58(9):1645–52.

Woermann, F. G., G. J. Barker, K. D. Birnie, H. J. Meencke, and J. S. Duncan. 1998. “Regional Changes in Hippocampal T2 Relaxation and Volume: A Quantitative Magnetic Resonance Imaging Study of Hippocampal Sclerosis.” *Journal of Neurology, Neurosurgery, and Psychiatry* 65(5):656–64.

Wolz, Robin, Paul Aljabar, Joseph V Hajnal, Alexander Hammers, Daniel Rueckert, and the

Alzheimer's Disease Neuroimaging Alzheimer's Disease Neuroimaging Initiative. 2010. "LEAP: Learning Embeddings for Atlas Propagation." *NeuroImage* 49(2):1316–25.

Xing, Z., X. Yang, D. She, Y. Lin, Y. Zhang, and D. Cao. 2017. "Noninvasive Assessment of IDH Mutational Status in World Health Organization Grade II and III Astrocytomas Using DWI and DSC-PWI Combined with Conventional MR Imaging." *American Journal of Neuroradiology* 38(6):1138–44.

Yim, Younghee, Ji Young Lee, Se Won Oh, Mi Sun Chung, Ji Eun Park, Yeonsil Moon, Hong Jun Jeon, and Won-Jin Moon. 2021. "Comparison of Automated Brain Volume Measures by NeuroQuant vs. Freesurfer in Patients with Mild Cognitive Impairment: Effect of Slice Thickness." *Yonsei Medical Journal* 62(3):255.

Zhang, James Y., Brent D. Weinberg, Ranliang Hu, Amit Saindane, Mark Mullins, Jason Allen, and Michael J. Hoch. 2019. "Quantitative Improvement in Brain Tumor MRI Through Structured Reporting (BT-RADS)." *Academic Radiology* 0(0).

Zhang, Li, Zhiqian Min, Min Tang, Sipan Chen, Xiaoyan Lei, and Xiaoling Zhang. 2017. "The Utility of Diffusion MRI with Quantitative ADC Measurements for Differentiating High-Grade from Low-Grade Cerebral Gliomas: Evidence from a Meta-Analysis." *Journal of the Neurological Sciences* 373:9–15.

Zhou, Hao, Martin Vallières, Harrison X. Bai, Chang Su, Haiyun Tang, Derek Oldridge, Zishu Zhang, Bo Xiao, Weihua Liao, Yongguang Tao, Jianhua Zhou, Paul Zhang, and Li Yang. 2017. "MRI Features Predict Survival and Molecular Markers in Diffuse Lower-Grade Gliomas." *Neuro-Oncology* 19(6):862–70.

Zivadinov, R., N. Bergsland, J. R. Korn, M. G. Dwyer, N. Khan, J. Medin, J. C. Price, B. Weinstock-Guttman, D. Silva, D. P. Ramasamy, E. Carl, S. Hunter, T. Scott, K. Pandey, E. Fox, A. Katz, J. Silversteen, J. Kaplan, E. Maa, V. Simnad, R. Shin, S. Newman, P. Kinkel, B. Green, J. Calkwood, K. Edwards, D. Jacobs, D. Huang, A. Bass, S. Hibbs, I. Kister, G. Eubank, G. Pardo, C. Chitnis, B. Hendin, S. Cohan, M. Freedman, M. Gudesblatt, E. Lathi, E. Alvarez, A. Goodman, R. Trudell, and R. Naismith. 2018. "Feasibility of Brain Atrophy Measurement in Clinical Routine without Prior Standardization of the MRI Protocol: RESULTS from Ms-Mrius, a Longitudinal Observational, Multicenter Real-World Outcome Study in Patients with Relapsing-Remitting MS." *American Journal of Neuroradiology* 39(2):289–95.