Machine Learning And Quantitative Neuroimaging In Epilepsy And Low Field MRI

Thomas Campbell Arnold
University of Pennsylvania

Follow this and additional works at: https://repository.upenn.edu/edissertations

Part of the Medicine and Health Sciences Commons

Recommended Citation
Arnold, Thomas Campbell, "Machine Learning And Quantitative Neuroimaging In Epilepsy And Low Field MRI" (2022). Publicly Accessible Penn Dissertations. 5355.
https://repository.upenn.edu/edissertations/5355

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/edissertations/5355
For more information, please contact repository@pobox.upenn.edu.
Abstract
Medical imaging plays a key role in the diagnosis and management of neurological disorders. Magnetic resonance imaging (MRI) has proven particularly useful, as it produces high resolution images with excellent tissue contrast, permitting clinicians to identify lesions and select appropriate treatments. However, demand for MRI services has outpaced the availability of qualified experts to operate, maintain, and interpret images from these devices. Radiologists often rely on time-consuming manual analyses, which further limits throughput. Moreover, a large portion of the world’s population cannot currently access MRI, and demand for medical imaging services will continue to increase as healthcare quality improves globally. To address these challenges, we must find innovative ways to automate medical processing and produce lower-cost medical imaging devices. Recent advances in deep learning and low-field MRI hardware offer potential solutions, providing lower-cost methods for processing and collecting images, respectively. This thesis aims to develop and validate lower-cost methods for collecting and interpreting neuroimaging using machine learning algorithms and portable, low-field MRI technology. In the first section, I develop a deep learning algorithm that automatically segments resection cavities in epilepsy surgery patients and quantifies removed tissues. I also compare the impacts of epilepsy surgery on remote brain regions, demonstrating that more selective procedures minimize postoperative cortical thinning. In the second section, I explore and validate clinical applications for a new portable, low-field MRI device. Using open-source imaging and machine learning, I propose a low-cost method for simulating diagnostic performance for novel imaging devices when only sparse data is available. Additionally, I validate device performance in multiple sclerosis by directly comparing the low-field device to standard-of-care imaging using a range of manual and automated analyses. My hope is that machine learning and low-field MRI will increase medical imaging access and improve patient care worldwide.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Bioengineering

First Advisor
Brian Litt

Keywords
healthcare cost, low-field MRI, MRI accessibility, multiple sclerosis, point-of-care MRI, portable MRI

Subject Categories
Medicine and Health Sciences

This dissertation is available at ScholarlyCommons: https://repository.upenn.edu/edissertations/5355
MACHINE LEARNING AND QUANTITATIVE NEUROIMAGING IN EPILEPSY AND
LOW FIELD MRI

Thomas Campbell Arnold

A DISSERTATION

in

Bioengineering

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2022

Supervisor of Dissertation

Brian Litt, M.D., Professor of Bioengineering and Neurology

Co-Supervisor of Dissertation

Dani S. Bassett, Ph.D., Professor of Bioengineering, Physics and Astronomy, Electrical and Systems Engineering, Neurology, and Psychiatry

Graduate Group Chairperson

Yale E. Cohen, Ph.D., Professor of Otorhinolaryngology

Dissertation Committee

Alison M. Pouch, Ph.D., Assistant Professor of Radiology
Kathryn A. Davis, M.D., Associate Professor of Neurology
Joel M. Stein, M.D., Ph.D., Assistant Professor of Radiology
MACHINE LEARNING AND QUANTITATIVE NEUROIMAGING IN EPILEPSY AND
LOW FIELD MRI
COPYRIGHT
2022
Thomas Campbell Arnold

This work is licensed under the Creative Commons
Attribution-NonCommercial-ShareAlike 4.0 License.

To view a copy of this license, visit
https://creativecommons.org/licenses/by-nc-sa/4.0/us/
To Agnes,

I’ll get a job eventually.
ACKNOWLEDGEMENT

This thesis was only possible through the unwavering support of my mentors, family, friends, colleagues, and collaborators. If you hate it, blame them.

I would like to start by thanking my primary advisor, Brian Litt. The lab environment you have created is uniquely both welcoming and ambitious, and I am honored to be a part of it. When students visit, I always describe you as a mentor that “lets you steer your own ship”. The task of self-definition has been the most challenging and rewarding aspect of my PhD, and I am eternally grateful for the growth I’ve experienced under your guidance. I would also like to thank my co-advisors, Kate Davis and Joel Stein, who have provided guidance throughout my PhD. They put in countless hours meeting with me to refine my projects and professional thinking. I would also like to thank Dani Bassett for teaching me to seek the question, not the solution.

I want to acknowledge the HHMI program directors, including Jim Gee, Alison Pouch, and Andrew Maidment for giving me the opportunity to be embedded in the medical school. I learned an immense amount during this time, and it has directly impacted my career trajectory.

I would also like to thank my past mentors, Wen Li and the larger-than-life Dennis Slice. Your belief in a naïve undergraduate allowed me to attend the University of Pennsylvania.

I knew I wanted to be a part of the Center for Neuroengineering and Therapeutics within the first 15 minutes of being there. The people were friendly, the students seemed happy, and it felt like home. I would like to thank the CNT staff, especially Carolyn Wilkinson, Adam Gibson, and Jacqueline Boccanfuso, who have generously given their time to help me with grant submissions and data management. To all my lab mates and colleagues, you have made my PhD a wonderful, fulfilling experience. I would especially like to recognize my cohort, Andy, Brittany, and John, your comradery, support, and memes have been essential to my survival.
My research has been supported by amazing collaborators from academia and industry. To Taki Shinohara, your unparalleled network centrality and statistical deep-dives have been invaluable. I would like to thank Danny Reich and his team at the NIH as well as Vicky Morgan and Dario Englot at VUMC for being excellent collaborators. I am grateful to the researchers at Hyperfine, including Samantha By, Brian Welch, Harry Hu, and John Pitts, for giving me the opportunity to work on this amazing technology. I have also been supported by numerous Penn faculty including Sandy Das, Phil Cook, Colbey Freeman, and Matt Schindler.

I would also like to recognize work done by the Bioengineering Department staff, especially Kate Venit, to run this amazing program. I am honored to have been a part of the Graduate Association of Bioengineers, which has established a community in our department and fostered numerous friendships. Additionally, I owe no small amount of gratitude to the Neuroradiology Research Core, for supporting my research and helping to implement low-field MRI at the Hospital of the University of Pennsylvania.

To the students I’ve mentored, especially Ramya Muthukrishnan, I am prouder of your accomplishments than any paper I’ve published.

To my parents and my extended family, you have impacted me in ways that I cannot articulate. Finally, to my wife, Agnes Isabel Crespo Arnold, thank you for inspiring and supporting my dreams. I can’t wait for our next adventure together.
ABSTRACT

MACHINE LEARNING AND QUANTITATIVE NEOUROIMAGING IN EPILEPSY
AND LOW FIELD MRI

Thomas Campbell Arnold
Brian Litt

Medical imaging plays a key role in the diagnosis and management of neurological disorders. Magnetic resonance imaging (MRI) has proven particularly useful, as it produces high resolution images with excellent tissue contrast, permitting clinicians to identify lesions and select appropriate treatments. However, demand for MRI services has outpaced the availability of qualified experts to operate, maintain, and interpret images from these devices. Radiologists often rely on time-consuming manual analyses, which further limits throughput. Moreover, a large portion of the world’s population cannot currently access MRI, and demand for medical imaging services will continue to increase as healthcare quality improves globally. To address these challenges, we must find innovative ways to automate medical processing and produce lower-cost medical imaging devices. Recent advances in deep learning and low-field MRI hardware offer potential solutions, providing lower-cost methods for processing and collecting images, respectively. This thesis aims to develop and validate lower-cost methods for collecting and interpreting neuroimaging using machine learning algorithms and portable, low-field MRI technology. In the first section, I develop a deep learning algorithm that automatically segments resection cavities in epilepsy surgery patients and quantifies removed tissues. I also compare the impacts of epilepsy surgery on remote brain regions, demonstrating that more selective procedures minimize postoperative cortical thinning. In the second section, I explore and validate clinical applications for a new portable, low-field MRI device. Using open-source imaging and machine learning, I propose a low-cost method for simulating diagnostic performance for novel imaging devices when only sparse data is
available. Additionally, I validate device performance in multiple sclerosis by directly comparing the low-field device to standard-of-care imaging using a range of manual and automated analyses.

My hope is that machine learning and low-field MRI will increase medical imaging access and improve patient care worldwide.
# TABLE OF CONTENTS

**ACKNOWLEDGEMENT** .......................................................................................................................... iv

**ABSTRACT** ............................................................................................................................................... vi

**LIST OF TABLES** ................................................................................................................................. xi

**LIST OF ILLUSTRATIONS** .................................................................................................................. xii

Chapter 1 - *Introduction* ......................................................................................................................... 1

Chapter 2 - *Automated Segmentation of Epilepsy Surgical Resections* .............................................. 4

Abstract ....................................................................................................................................................... 5

2.1 Introduction ............................................................................................................................................. 6

2.2 Materials and Methods .......................................................................................................................... 8
  2.2.1 Data collection ........................................................................................................................................ 8
  2.2.2 Data preprocessing ................................................................................................................................. 8
  2.2.3 Model Architecture ............................................................................................................................... 8
  2.2.4 Model Training ...................................................................................................................................... 9
  2.2.5 Post processing .................................................................................................................................... 9
  2.2.6 Model Evaluation ................................................................................................................................. 9
  2.2.7 Quantifying surgical remnants ............................................................................................................. 11
  2.2.8 Code availability ................................................................................................................................. 12

2.3 Results .................................................................................................................................................. 12
  2.3.1 Patient characteristics ........................................................................................................................... 12
  2.3.2 Primary performance measure (DSC) ................................................................................................. 14
  2.3.3 Secondary performance measures (Hausdorff distance & volume) ..................................................... 14
  2.3.4 False negatives & false positives ...................................................................................................... 16
  2.3.5 Lesion size relationships .................................................................................................................... 16
  2.3.6 Comparing surgical approaches ...................................................................................................... 17
  2.3.7 Visual inspection .................................................................................................................................. 17
  2.3.8 Quantifying surgical remnants ........................................................................................................... 18

2.4 Discussion ............................................................................................................................................ 20

2.5 Conclusion .......................................................................................................................................... 24

Chapter 3 - *Remote effects of temporal lobe epilepsy surgery* ........................................................... 25

3.1 Introduction .......................................................................................................................................... 28

3.2 Methods ............................................................................................................................................... 30
  3.2.1 Patients .............................................................................................................................................. 30
  3.2.2 Image Acquisition ............................................................................................................................. 31
  3.2.3 Selection of surgical approach ......................................................................................................... 32
  3.2.4 Resection Segmentation ................................................................................................................... 32
  3.2.5 Cortical Thinning Estimation .......................................................................................................... 33
  3.2.6 Statistical Analysis ............................................................................................................................ 35

3.3 Results ................................................................................................................................................. 36
  3.3.1 Demographics ................................................................................................................................. 36
LIST OF TABLES

Table 1 - Patient demographics information ................................................................. 13
Table 2 - Cross-validation results .................................................................................... 14
Table 3 - Demographics of patients included in study .................................................. 31
Table 4 – List of open-access neuropathology datasets used in this study .................. 86
Table 5 – Performance metrics for each pathology type using standard or simulated low-field images ........................................................................................................ 95
Table 6 – Patient Demographics .................................................................................... 111
Table 7 – Sequence parameters for study scans ............................................................ 112
LIST OF ILLUSTRATIONS

Figure 1 - Automated Segmentation of Epilepsy Surgical Resections ........................................ 10
Figure 2 - Pipelines for automated resection segmentation and quantification of postsurgical volume estimates ................................................................................................................................. 12
Figure 3 - Classifier accuracy across patient cohort ..................................................................... 15
Figure 4 - Example output from the lowest scoring segmentations .............................................. 18
Figure 5 - Graphical User Interface (GUI) for estimating surgical remnants ............................. 19
Figure 6 - Strong correlation between remnant estimates using automated and manual methods ................................................................................................................................. 20
Figure 7 - Tissue deformation errors ............................................................................................ 34
Figure 8 - Analytical pipeline ........................................................................................................ 35
Figure 9 - Preoperative cortical thickness ..................................................................................... 38
Figure 10 - Probability maps for each surgical group ................................................................. 39
Figure 11 - Changes in postoperative cortical thickness .............................................................. 40
Figure 12 - Differences in cortical thinning rates between ATL and SAH .................................. 41
Figure 13 - Correlations between clinical variables and cortical thinning ................................ 42
Figure 14 - Defining “low-field” MRI .......................................................................................... 50
Figure 15 - Research interest in low-field ....................................................................................... 52
Figure 16 - Examples of hemorrhage and Midline shift (MLS) on a 64mT scanner ..................... 58
Figure 17 - Images from a 27-year-old patient treated with ferumoxytol for iron deficiency anemia ............................................................................................................................. 60
Figure 18 - Patients imaged with 3T (top) and 64mT (bottom) at our center .............................. 63
Figure 19 - Generating simulated low-field (LF) images ............................................................. 87
Figure 20 - Validation of image transformation method .............................................................. 89
Figure 22 - Radiologist ratings of real and simulated images ...................................................... 94
Figure 23 - Pathology detection performance ............................................................................. 95
Figure 24 - Detection sensitivity as a function of lesion size and scaled intensity .................... 96
Figure 25 - Model validation and interpretability ........................................................................ 98
Figure 26 - Intensity modulation to explore effects of intensity contrast ................................. 99
Figure 27 - Flow chart of study participants ................................................................................. 109
Figure 28 - MS lesions appear similar on 3T and 64mT pulse sequences .................................... 110
Figure 29 - Automated segmentation algorithm ......................................................................... 113
Figure 30 - Manual lesion size measurements and interrater reliability ................................ 118
Figure 31 - Automated lesion segmentations at 3T and 64mT overlap .................................... 120
Figure 32 - Total lesion volume measured at 3T and 64mT shows agreement ........................... 121
Figure 33 - 64mT image artifacts ................................................................................................. 122
Figure 34 - False positive lesion segmentations on 64mT T2-FLAIR ......................................... 123
Figure 35 - Dice score increases with lesion volume .................................................................... 124
Figure 36 - Lesion size and intensity influence detection rate ...................................................... 125
Figure 37 - Multi acquisition image averaging increases resolution .......................................... 126
Medical imaging is an essential tool in modern healthcare. This technology assists clinicians in a myriad of tasks, including diagnosis, tracking disease progression, monitoring treatments for efficacy or complications, surgical planning, and preventative care through screening. In fact, the discovery of X-rays was deemed so revolutionary that it garnered Wilhelm Roentgen the first ever Noble prize in Physics. Since then, 25 Nobel prizes have been awarded for medical imaging discoveries, with at least 13 related to Magnetic Resonance Imaging (MRI) (1,2). MRI is particularly useful because it produces high-resolution, cross-sectional images with excellent soft tissue contrast, permitting clinicians to see internal organs, tissues, and the skeletal system in exquisite detail.

MRI has proven to be invaluable in the diagnosis and management of neurological disorders, which affect over one billion people worldwide (3). Neurological disorders are the leading cause of disability and the second leading cause of death, killing an estimated 9 million people annually (4). MRI is particularly relevant for the two primary diseases considered in this thesis, multiple sclerosis and epilepsy. In multiple sclerosis (MS), key MRI features in the McDonald criteria establish the diagnosis, determine the treatment course, and indicate clinical trial eligibility (5,6). In epilepsy, identification of a lesion on neuroimaging is the leading predictor of postoperative seizure freedom (7). Additionally, MRI offers an alternative to costly and invasive intracranial procedures for localizing the epileptic focus for surgical guidance (8).

Despite the apparent benefits of MRI, it remains underutilized and unavailable in a large portion of the world. Approximately ninety percent of people do not have access to MRI (9), and an astounding two-thirds of the world’s population does not have access to even basic medical imaging (10–12). Furthermore, even where medical imaging is available many neuroimaging
applications have yet to be widely translated into clinical practice (13). Despite the known power of imaging, it remains too costly or impractical to scan patients with high frequency or perform tedious manual analyses. The leading causes for limited neuroimaging utilization are 1) high device costs (i.e. purchase, maintenance, and operation), 2) lack of training (i.e. in operation and image interpretation), and 3) high burden of use (i.e. labor intensive or clinically infeasible analyses) (9,13).

To improve MRI access for patients with neurologic disorders, this research aims to develop machine learning tools that automate neuroimaging analysis and to evaluate the diagnostic capabilities of a lower-cost, portable low-field MRI. Recent advances in machine learning have resulted in significant improvements in automated image interpretations and identification of key structures (14). These advancements can be leveraged to reduce the burden of intensive manual analyses and perhaps even permit image interpretation in situations where a trained radiologist or radiographer is not readily available (15). Additionally, new lower-cost MRI devices have become available as a result of hardware and software advancements. Machine learning and low-field MRI, in combination, could broaden the reach of MRI and significantly impact how patients are treated for neurological disorders across the globe.

This thesis focuses on studies that evaluate lower cost options for clinical neuroimaging acquisition and analysis, thus broadening access to the high-quality medical imaging necessary to treat neurological disorders. I will focus on two overarching applications: (i) automated analysis of epilepsy neuroimaging, and (ii) evaluation of portable low-field MRI for detecting neuropathology. The thesis is comprised of seven chapters relating to these broader topics.

The next two chapters are centered on the development of automated tools for analyzing postoperative epilepsy neuroimaging. Chapter 2 addresses a key challenge when working with patients that have undergone epilepsy surgery, the accurate identification of the resection cavity. Precise delineation of the resection cavity is essential for automated postoperative imaging
processing and for researchers to investigate why certain patients do not achieve seizure freedom after surgery. This chapter describes the development of an automated algorithm that uses deep learning to provide robust resection segmentations. Chapter 3 describes the impacts of epilepsy surgery on downstream brain regions. Patients that have undergone epilepsy surgery are understudied in part because processing postoperative imaging with conventional image analysis pipelines is difficult. Using the automated method for resection segmentation, we can precisely delineate the surgical boundaries and adapt existing image processing pipelines to account for this region. These tools are used to compare two common approaches to epilepsy surgery and describe differences in how these approaches affect the brain after surgery.

The subsequent section focuses on clinical validation of a lower-cost, portable MRI. In Chapter 4 we review recent advances in low-field MRI technology and promising clinical applications. Chapter 5 describes a computational simulation that can be used to evaluate proposed low-field devices across a range of pathologies. This method utilizes open-access data and machine learning to generate and interpret simulated low-field images. The simulation approach can be used to probe potential applications and provide guidance for prospective studies. Chapter 6 presents a validation of the low-field MRI device in a cohort of patients with multiple sclerosis. Paired low-field and high-field imaging are collected and directly compared using both manual and automated interpretation methods. This work provides evidence that low-field MRI can be used to assess multiple sclerosis and highlights key challenges that must be addressed for these devices to be used clinically. Finally, the thesis concludes with a discussion of key takeaways and potential future work that can further improve access to medical imaging.
Accurate segmentation of surgical resections is crucial for clinical assessments and neuroimaging research applications in epilepsy, including predicting surgical outcomes. In this chapter, we present an automated resection segmentation algorithm that was developed using manual resection segmentations of postoperative temporal lobe epilepsy patients. We deployed our algorithm in an easy-to-use graphical user interface (GUI) that estimates resected brain regions. This work is currently under review at NeuroImage: Clinical, with T. Campbell Arnold as co-first author.

Deep Learning-Based Automated Segmentation of Resection Cavities on Postsurgical Epilepsy MRI

T. Campbell Arnold\textsuperscript{1,2}, Ramya Muthukrishnan\textsuperscript{2,3}, Akash R. Pattnaik\textsuperscript{1,2}, Adam Gibson\textsuperscript{2}, Nishant Sinha\textsuperscript{2}, Sandhitsu R. Das\textsuperscript{4}, Brian Litt\textsuperscript{1,2,5}, Dario J. Englot\textsuperscript{6-9}, Victoria L. Morgan\textsuperscript{7-9}, Kathryn A. Davis\textsuperscript{2,5}, Joel M. Stein\textsuperscript{4}

1. Department of Bioengineering, University of Pennsylvania
2. Center for Neuroengineering and Therapeutics, University of Pennsylvania
3. Department of Computer Science, University of Pennsylvania
4. Department of Radiology, University of Pennsylvania
5. Department of Neurology, University of Pennsylvania
6. Department of Neurological Surgery, Vanderbilt University Medical Center
7. Department of Radiology, Vanderbilt University Medical Center, Nashville
8. Department of Biomedical Engineering, Vanderbilt University Medical Center
9. Institute of Imaging Science, Vanderbilt University Medical Center

* These authors contributed equally to this work.
Abstract

Accurate segmentation of surgical resection sites is critical for clinical assessments and neuroimaging research applications, including resection extent determination, predictive modeling of surgery outcome, and masking image processing near resection sites. In this study, an automated resection cavity segmentation algorithm is developed for analyzing postoperative MRI of epilepsy patients and deployed in an easy-to-use graphical user interface (GUI) that estimates resected brain regions. This retrospective study included postoperative T1-weighted MRI from 45 temporal lobe epilepsy (TLE) patients who underwent resective surgery. The resection site was manually segmented and reviewed by a neuroradiologist (JMS). A U-Net convolutional neural network was trained to segment surgical resection in axial slices. The algorithm was trained using 5-fold cross validation, with data partitioned into training (N = 27), testing (N = 9), and validation (N = 9) sets. Algorithm performance was assessed using Dice-Sørensen coefficient (DSC), Hausdorff distance, and volume estimates. Additionally, we deploy a fully-automated, GUI-based pipeline that compares resection segmentations with preoperative imaging and reports estimates of resected brain structures. The median DSC between manual and automated segmentations was 0.80 (0.12 interquartile range), which approaches inter-rater reliability between radiologists (0.84 - 0.86) as reported in the literature. Median 50% and 95% Hausdorff distance was 1.0 mm and 4.1 mm respectively, indicating high segmentation boundary confidence. Automated and manual resection volume estimates were highly correlated (r = 0.85, p < 0.0001, mean average error = 5.2%). Automated and manual segmentations overlapped in all 45 subjects, indicating a low false negative rate. In control subjects (n = 5), no resection was detected (i.e. 0% false positives). There was strong agreement between postoperative hippocampal remnant volumes made using automated and manual resection segmentations (r = 0.88, p < 0.0001, mean average error = 7.6%), indicating that automated resection segmentations can permit quantification of postoperative brain volumes after...
epilepsy surgery. Applications include quantification of postoperative remnant brain volumes, correction of deformable registration, and localization of removed brain regions for network modeling.

### 2.1 Introduction

Epilepsy is a neurological disorder characterized by recurrent seizures, affecting sixty-five million people worldwide (16). Temporal lobe epilepsy (TLE) is the most common form of epilepsy, with a prevalence of 8.9 cases per 100,000 people per year in the US (17). Surgical removal of the epileptic focus is the recommended treatment for drug-resistant TLE, however only about 60% of patients experience seizure-freedom one year postoperatively (18–22). A variety of approaches have been used to better predict surgical outcome, including quantitative assessments of resection extent (23–25) and modelling the surgical procedure on preoperative functional or structural networks (25,26). However, many of these methods rely on manual segmentation of the resection zone and automated methods for quantifying resection extent would be of significant interest to the epilepsy clinical and neuroimaging communities.

Retrospective studies attempting to predict surgical outcome of TLE patients use a variety of manual methods which are susceptible to bias due to inter-rater variability. Simple measures, such as which brain regions fall within the resection zone, have been identified as a positive predictor of postoperative seizure freedom (24,27). More complex methods mimic surgical resection on brain network models to predict postoperative seizure freedom (28,29). The resected brain regions are often determined through manual segmentation or visual inspection (26,30). Time-consuming and error-prone manual methods for determining resected tissue limit clinical adoption of these tools. An automated method for delineating resection cavities would be of
substantial clinical and research interest, with potential to increase the accuracy of predictive models, evaluate alternative surgical strategies, and improve patient outcomes.

Recent advances in convolutional neural networks (CNNs) have led to vast improvements in classification and segmentation of medical imaging (31). Neural network architectures designed for segmentation, such as U-Net, have been successfully applied to problems from a wide range of specialties, including radiology, pathology, and dermatology (14). Additionally, new deep-learning enabled neuroimaging software packages have dramatically reduced processing time for tasks such as brain parcellation (32). A primary goal for many of these tools is the automation of tedious, time-consuming tasks in medicine (33). In epilepsy patient care, predictive models of resection extent have not been adopted clinically in part due to their reliance on manual segmentation of resections or visual inspection by researchers, which is time-consuming and variable across individuals and institutions. The lack of automated resection segmentation methods prevents quantitative neuroimaging analyses from being integrated into epilepsy patient care (34).

Therefore, the goal of our study was to develop a fully automated method for segmenting resection volumes and quantifying resected brain tissues in TLE patients. This tool can evaluate successful removal of surgical targets and has potential to improve predictive models of surgical outcome. Additionally, we present a graphical user interface (GUI) that allows users who are not familiar with machine learning to easily apply the model to their data. We demonstrate that the model can segment the resection zone and estimate which brain regions were removed in under 5 minutes, permitting easy integration into a clinical workflow. We openly share all code, including the machine learning model, GUI, and statistical analyses to facilitate clinical translation of our work.
2.2 Materials and Methods

2.2.1 Data collection

T1-weighted images (N = 45) were collected from temporal lobe epilepsy (TLE) patients who underwent surgery at the Hospital of the University of Pennsylvania (HUP) or Vanderbilt University Medical Center (VUMC). Internal Review Boards of each institution approved this study, and all patients gave informed consent. Patients at HUP (N = 22) were imaged primarily using a Siemens 3T scanner with the following T1-weighted sequence parameters: 1mm isotropic, TE = 3.87ms, TR = 1.62s, and flip-angle = 15. Patients at VUMC (N = 23) were imaged using a Phillips 3T scanner and T1-weighted sequence parameters were 1mm isotropic, TE = 4.61ms, TR = 8.9ms, and flip-angle = 8. All images were collected at least 5 months postoperatively to avoid peri-surgical swelling. Inclusion criteria were: 1) TLE patients that underwent resection, 2) whole-brain, isotropic T1-weighted imaging at least 5 months postoperatively, and 3) only one contiguous resection site. Preoperative T1-weighted imaging using the same sequence parameters were available for 36 of the 45 patients. Additionally, T1-weighted images were collected from five healthy participants at HUP using the same protocol.

2.2.2 Data preprocessing

The resection site in each postoperative T1-weighted image was manually segmented in ITK-SNAP (35) and reviewed by a board certified neuroradiologist with 8 years of experience (JMS). Each 3D volumetric image was normalized to a standard intensity range [0-1], and 2D axial slices were output as 256x256 Portable Networks Graphic (PNG) files for training.

2.2.3 Model Architecture

A U-Net CNN architecture was used to segment resections (14). Model construction and training was carried out using the Keras API with TensorFlow backend (36). The model training
script was adapted from an open-source U-net segmentation project (37) to our model architecture and run on an independent server using a Titan-X GPU. The U-Net architecture consists of an encoder that captures contextual information and a decoder that captures localization information to output a predicted mask. Our model architecture replaced the traditional U-Net encoder with the EfficientNet B1 network encoder backbone, and initial encoder weights were pre-trained on ImageNet data (38,39).

2.2.4 Model Training

The model performed binary segmentation of resections (i.e. 1 = resected, 0 = not resected) in axial slices of T1-weighted images. During model training, 5-fold cross validation was employed with data divided into training, validation, and test sets (3:1:1 split). All axial slices for a given subject were contained within a single set (i.e. training, validation, or testing). The 5-fold approach permits each subject in the dataset to be included in the test set once. Models were trained for 50 epochs using the Adam optimizer, a learning rate of 1e-4, and a batch size of 16. Data augmentation was employed during training to increase model generalizability (40). Augmentation included random horizontal and vertical flips, rotations up to 10 degrees, and horizontal and vertical shifts up to 10% of image width and height.

2.2.5 Post processing

Segmentations output by the CNN underwent two post processing steps: 1) assembly of 3D volumes from 2D slice segmentations and 2) connected components analysis to remove isolated voxels.

2.2.6 Model Evaluation

Segmentation performance was primarily evaluated using the Dice-Sørensen coefficient (DSC) (41). DSC measures the overlap between manual segmentations $X$ and automated
segmentations $Y$ (Figure 1), by computing: $DSC = \frac{2|X \cap Y|}{|X| + |Y|}$. Performance for DSC ranges from 0 (no overlap) to 1 (perfect match). Hyperparameter optimization was driven by DSC maximization in the validation set. To assess model generalizability to novel images, all reported DSC values were calculated on held-out test datasets.

Subtle changes to performance measures can result in significant differences when ranking algorithms (42). To provide a holistic view of model performance, we report multiple metrics and descriptive statistics. We report two secondary measures, Hausdorff distance and total resection volume (Figure 1). Hausdorff distance compares actual and predicted segmentation boundaries and reports the distances between adjacent boundary points. This measure characterizes the segmentation border reliability. Several variants of the Hausdorff distance are reported in the literature (43,44). For comparison we report 50% and 95% Hausdorff distances. Additionally, the
relationship between manual and automated resection volumes was plotted for each subject and we report the Pearson’s correlation coefficient and mean absolute error (MAE) between these variables.

2.2.7 Quantifying surgical remnants

Postoperative remnant volumes, such as the hippocampal remnant, have predictive value for TLE surgical outcome (27). We further developed a pipeline that estimates postoperative remnant brain structures. The pipeline takes a patient’s preoperative and postoperative T1 brain MRI as input and generates a PDF report or an interactive web-based report of estimated resection impact on brain structures (Figure 2). Preoperative imaging was coregistered to post-operative imaging and segmented into brain regions using the Desikan–Killiany–Tourville (DKT) atlas with subcortical parcellations using the deep-learning enabled toolkit, Advanced Normalization Tools Python (ANTsPyNet) (32,45,46). Images were coregistered using a symmetric normalization transformation, with cross-correlation as optimization metric and cost-function masking of the resection zone to mitigate image distortion (47). Proper image coregistration was verified manually and any subjects with significant distortion (N=1) were excluded. The resection cavity was segmented both manually and using the automated algorithm described here for comparison. The intersection of preoperative brain segmentations and the postoperative resection segmentation was used to estimate remnant brain volumes. We correlated hippocampal remnant volume estimates between manual and automated resection segmentation methods.
Figure 2 - Pipelines for automated resection segmentation and quantification of postsurgical volume estimates. The resection segmentation pipeline uses a U-Net architecture (top) and produces a 3D binary mask of resected tissue. To quantify postoperative remnant volumes (bottom), the preoperative image was segmented into brain regions. The intersection of the resection and functional brain segmentations were used to generate a resection report.

2.2.8 Code availability

All code related to model design and postoperative volume estimation can be found at: https://github.com/penn-cnt/DeepResection. Statistical analyses are available in a Jupyter-Notebook at: https://github.com/penn-cnt/DeepResection_Statistical_Analysis.

2.3 Results

2.3.1 Patient characteristics

Our dataset included 45 patients who underwent surgery for localization-related epilepsy across two institutions, Hospital of the University of Pennsylvania (HUP, N = 22) and Vanderbilt University Medical Center (VUMC, N = 23). Patients were age-matched across institution (HUP:
39.2 ± 12.0 years, VUMC: 39.5 ± 12.7 years) and were treated with either anterior temporal lobectomy (ATL, N=28), selective amygdalohippocampectomy (SAH, N=15), or hippocampal laser interstitial thermal therapy (LITT, N=2). There was no significant difference between institutions for the patients’ gender, age of seizure onset, age at surgery, side of seizure surgery, age at scan, or disease duration. There was a significant difference in the surgical approach between institutions ($X^2 = 21.99, p < 0.0001, \text{chi-square test}$), as patients were only treated with SAH at one center. Demographic information is available in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>VUMC</th>
<th>HUP</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex (female / male)</td>
<td>12 / 11</td>
<td>16 / 6</td>
<td>28 / 17</td>
</tr>
<tr>
<td>Surgical approach</td>
<td>7 / 15 / 1</td>
<td>21 / 0 / 1</td>
<td>28 / 15 / 2</td>
</tr>
<tr>
<td>(ATL / SAH / LITT)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Side of surgery</td>
<td>6 / 17</td>
<td>11 / 11</td>
<td>17 / 28</td>
</tr>
<tr>
<td>(left / right)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age at surgery</td>
<td>37.5 ± 11.5</td>
<td>36.6 ± 12.8</td>
<td>37.1 ± 12.1</td>
</tr>
<tr>
<td>(years, mean ± SD)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age at scan</td>
<td>39.2 ± 12.0</td>
<td>39.5 ± 12.7</td>
<td>39.3 ± 12.2</td>
</tr>
<tr>
<td>(years, mean ± SD)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age at onset</td>
<td>19.5 ± 10.6</td>
<td>18.9 ± 12.2</td>
<td>19.1 ± 11.6</td>
</tr>
<tr>
<td>(years, mean ± SD)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration</td>
<td>15.6 ± 8.7</td>
<td>21.8 ± 15.7</td>
<td>19.9 ± 14.1</td>
</tr>
<tr>
<td>(years, mean ± SD)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 1 - Patient demographics information.** Abbreviations: Hospital of the University of Pennsylvania (HUP), Vanderbilt University Medical Center (VUMC), anterior temporal lobectomy (ATL), selective amygdalohippocampectomy (SAH), Laser interstitial thermal therapy (LITT), standard deviation (SD).
2.3.2 Primary performance measure (DSC)

The model was trained using 5-fold cross validation and accuracy is reported using the per-scan DSC for held out subjects (Table 2). The average test DSC across all scans was 0.77 ± 0.09 (mean ± standard deviation), with a median DSC of 0.80 and interquartile range of 0.12 (Figure 3A). The maximum DSC achieved by the classifier for a given patient was 0.89, while the minimum score was 0.49. To illustrate the range of segmentation quality, Figure 3D visualizes the manual segmentation and corresponding predicted label at each quartile of the DSC distribution.

<table>
<thead>
<tr>
<th>Fold</th>
<th>DSC Median / IQR</th>
<th>50% HD Median / IQR</th>
<th>95% HD Median / IQR</th>
<th>100% HD Median / IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.81 / 0.14</td>
<td>0.98 / 0.02</td>
<td>4.77 / 3.84</td>
<td>10.42 / 11.93</td>
</tr>
<tr>
<td>2</td>
<td>0.80 / 0.12</td>
<td>1.00 / 0.02</td>
<td>3.91 / 0.26</td>
<td>9.43 / 5.45</td>
</tr>
<tr>
<td>3</td>
<td>0.83 / 0.03</td>
<td>1.00 / 0.02</td>
<td>4.12 / 1.34</td>
<td>10.20 / 7.28</td>
</tr>
<tr>
<td>4</td>
<td>0.79 / 0.17</td>
<td>1.00 / 0.40</td>
<td>5.09 / 3.00</td>
<td>10.84 / 2.53</td>
</tr>
<tr>
<td>5</td>
<td>0.79 / 0.09</td>
<td>1.00 / 0.02</td>
<td>5.00 / 3.83</td>
<td>9.80 / 3.64</td>
</tr>
<tr>
<td>All</td>
<td>0.80 / 0.12</td>
<td>1.00 / 0.02</td>
<td>4.12 / 3.34</td>
<td>10.06 / 7.28</td>
</tr>
</tbody>
</table>

**Table 2** - Cross-validation results. Two metrics are reported, Dice-Sørensen coefficient (DSC) and Hausdorff distance (HD).

2.3.3 Secondary performance measures (Hausdorff distance & volume)

Two secondary performance measures were assessed, Hausdorff distance and predicted resection volume. Hausdorff distance quantifies the maximal distance between analogous boundary points in ground truth and predicted segmentations. In our analyses we have included the 50% Hausdorff distance, which quantifies the average boundary error in a segmentation, as well as the
95% Hausdorff distance, which has been utilized as an accuracy measure in prominent segmentation challenges (24). In our dataset the median average Hausdorff distance was $1.0 \pm 0.02$ mm, while the median 95% Hausdorff distance was $4.1 \pm 3.3$ mm (Figure 3B). This indicates that on average the segmentation was one voxel away from the target boundary and that 95% of boundary points were within 4.1 mm of the target.

Figure 3 - Classifier accuracy across the patient cohort. (A) Dice-Sørensen coefficient (DSC), $0.80 \pm 0.12$ (median ± interquartile range). (B) 95% Hausdorff distance, $4.12 \pm 3.34$ mm (median ± interquartile range). (C) Pearson correlation between predicted and manually segmented volumes ($r = 0.85$, $p < 0.0001$). (D) Representative manual and automated segmentations from each quartile of the DSC distribution. Segmentations are overlaid on the T1-weighted images, with their associated DSC on the right-hand side.
When comparing manual and automated segmentation volumes, predicted volumes were slightly smaller (19.8 ± 13.9 ml) than manually labeled segmentations (23.0 ± 17.7 ml), with a significant difference detected in a pairwise t-test (p = 0.026, t = -2.30). There was a strong correlation (Figure 3C) between the manual and automated volumes (r = 0.85, p < 0.0001, MAE = 5.2%).

2.3.4 False negatives & false positives

The classifier’s false negative and false positive rates of lesion detection were also assessed. A false negative was defined as the algorithm having no segmentation overlap with the manual segmentation. A false positive was defined as inappropriate segmentation in a control subject. For all resection patients, the classifier correctly lateralized their resection to the appropriate hemisphere and the predicted resection overlapped the ground truth label. This indicates a low false negative rate for lesion detection (effectively 0%). To evaluate the potential for false positives, we separately processed 5 healthy control subjects with no history of neurological disorders. In all 5 control subjects no resection segmentations were produced. Importantly, this indicates the classifier is sensitive to the presence of a resection, not simply localizing the temporal lobe and producing an average resection mask as output.

2.3.5 Lesion size relationships

Previous studies have found a relationship between lesion size and classifier accuracy as measured by DSC and percent volume difference (PVD) between predicted and manual segmentations (44). To understand whether lesion size contributed to classifier accuracy or PVD error, we partitioned subjects into small (N = 17) and large (N = 29) resection groups using the same threshold (17.92 ml) previously reported (25). In our model, we found no relationship between the DSC and lesion size (p = 0.19, t = 1.34, two-sample t-test) though a modest effect was observed when comparing PVD for small and large lesion groups (p = 0.055, t = -1.97, two-sample
t-test) which indicates a larger percent error for small lesions. Predicted and actual volumes for small lesions were not significantly different (mean volume: actual = 7.2 ml, predicted = 7.1 ml, p = 0.87, t = 0.16, two-sample t-test), however large lesions tended towards under-segmentation (mean volume: actual = 32.6 ml, predicted = 27.5 ml, p=0.025, t = 2.37, two-sample t-test).

2.3.6 Comparing surgical approaches

Next, we compared the algorithm’s performance between patients treated with SAH (N = 15) and ATL (N = 28) to determine if one surgical approach accounted for a greater degree of model error. Patients treated with LITT were excluded from statistical analysis, as only two patients were in this group. There was no significant difference in our primary performance measure, DSC (ATL: 0.78 ± 0.1, SAH: 0.76 ± 0.08, p = 0.44, t = 0.77). Patients treated with LITT had slightly lower DSC (0.72-0.73), possibly caused by the low number of training samples and hyperintense coagulative necrosis in ablation cavities (48). There was a significant difference in the 50% Hausdorff distance, with ATL having a larger boundary error than SAH (ATL: 1.22 ± 0.73 mm, SAH: 0.73 ± 0.44 mm, p = 0.03, t = 2.3). However, at the more stringent 95% and 100% Hausdorff distances there was no measurable difference. As expected, both the ground truth volumes (ATL: 30.1 ± 12.3, SAH: 12.5 19.8, p = 0.001, t = 3.5) and predicted volumes (ATL: 25.6 ± 6.9, SAH: 11.2 ± 17.7, p < 0.001, t = 3.7) were significantly larger for patients treated with ATL.

2.3.7 Visual inspection

The largest source of error in estimating resection volume was in patients with resections that extended outside of the temporal lobe, including regions of the parietal lobe. Extra-temporal resections were poorly represented in the dataset, with only two such cases. Resections in these patients corresponded to the two lowest DSC scores (0.49-0.50) and were characterized by under-segmentation (Figure 4). Similarly, there were only two examples of patients treated with LITT. Although LITT patients had reasonable DSC scores (0.72-0.73), the predicted segmentation was
smaller than the ground truth in both cases. On visual inspection, the center of the ablation cavity was filled with hyperintense coagulative necrosis (48), which was not segmented by the algorithm. Additionally, there were cases of over-segmentation in comparison to the ground-truth label, particularly for smaller lesions. Cases of over-segmentation included anterior extension into the sphenoid wing and adjacent cerebral spinal fluid surrounding the resection site (Figure 3D).

Figure 4 - Example output from the lowest scoring segmentations. Both patients (A & B) had extra-temporal resections and were the only such cases in the dataset. Automated segmentations in these patients were undersegmented posteriorly and superiorly, as the resection extended outside the bounds of a traditional anterior temporal lobectomy.

2.3.8 Quantifying surgical remnants

The web application that deploys our pipeline consists of a set of sequential web pages where users can upload pre- and post-operative MR images, visualize the automated segmentation, and save a report estimating resected brain regions to their local desktop. The landing screen lets the user choose between applying the full pipeline to their data using our automated segmentation algorithm or uploading a manually generated segmentation and visualizing the report. The full
pipeline consists of pre- and post-operative image registration (32), pre-operative segmentation using the DKT brain atlas with subcortical structures (46), and resection segmentation using our described model. The report page consists of a table listing affected brain regions, a 3D resection mask viewer, and optional user feedback (Figure 5). The report table provides the total resection volume and lists affected brain regions by percentage resected. An embedded 3D mask viewer allows the user to make a quality assessment of the predicted mask against the post-operative image. To assess the feasibility of deploying the full pipeline to users, we computed the time elapsed for running the web application. The average run time was 4 minutes and 19 seconds, and all run times were less than 5 minutes.

Figure 5 - Graphical User Interface (GUI) for estimating surgical remnants. Here we illustrate the GUI interface developed for estimating resection remnants on a selective amygdalohippocampectomy patient. (A) In the first panel, the user selects to run the full pipeline or run the analysis using a resection volume they generated. (B) The user then uploads the required images and selects their desired registration and segmentation parameters. (C) The pipeline outputs a table listing affected regions by percentage resected and provides an interactive visualization of the resection segmentation for manual review and quality control.
Hippocampal remnants have previously been associated with worse surgical outcomes (27). We compared hippocampal remnant volumes using the manual and predicted resection segmentations for a subset of 36 patients with available preoperative imaging. One subject was excluded due to poor image registration. **Figure 6** illustrates the correlation between hippocampal remnant estimates made using the manual labels and automated segmentation ($r = 0.88$, $p < 0.0001$, MAE = 7.6%).

**Figure 6** - Strong correlation between remnant estimates using automated and manual methods. We compared hippocampal remnant estimates using automated and manual resection segmentations. Automated and manual estimates are significantly correlated ($r = 0.88$, $P < 0.00001$) and have a mean absolute error of 7.6%.

### 2.4 Discussion

We present a deep-learning method to fully automate resection cavity segmentation in postoperative temporal lobe epilepsy patients. Fully automated segmentation provides significant time advantages over manual and semi-automated methods (14,35,49). Our method has several key advantages. First, we trained our model explicitly on TLE patients, who are most frequently operated on for drug-resistant epilepsy. Second, our resection labels are based on gold standard
clinical practice (manually segmented by a neuroradiologist with 8 years of experience and subspecialization in epilepsy imaging). Third, we included multi-site data from two epilepsy centers, thus demonstrating generalizability and potential for multi-center studies. Fourth, we provide a fast method for volumetric analysis of resected brain regions for post-hoc analysis. Fifth, we incorporate a graphical user interface (GUI) for easy interpretation of segmentation quality. We demonstrate the clinical utility of our algorithm by quantifying postoperative remnant structures, which have been shown to predict long-term surgical outcome.

One advantage of the present study is epilepsy patient data was used during model training. A previous study applied lesion_GNB, a stroke segmentation classifier, to the resection cavity segmentation problem in epilepsy patients (44,50). While lesion_GNB demonstrated some utility in segmenting resections (median DSC 0.58), our classifier achieved a 48% increase (median DSC 0.80) in accuracy. The discrepancy in classifier performance may be caused in part by differences in features between the pathologies, such as lesion intensity and surrounding edema. Resection cavity segmentation has also been attempted in glioblastoma multiforme (GBM) patients (51). Here the classifier was trained explicitly on GBM patient data and classifier performance (median DSC 0.83) was similar to trained radiation oncologists (median DSC 0.85). These studies highlight the importance of developing disease specific classifiers or applying transfer learning to fine-tune models for specific pathologies (52).

Other approaches to boost classifier performance include the incorporation of simulated training data. Pérez-Garcia et al. recently reported the development of EPISURG, a self-supervised resection segmentation classifier that uses exclusively simulated resection data (33). Their classifier achieved a median DSC of 0.805 using 2074 simulated resections, which surpassed their classifier trained using 133 manual labels (median DSC 0.653). This illustrates the significant performance gains possible through innovative data augmentation. However, it is important to note that all
versions of their model report false negatives, meaning in some subjects the resection was entirely missed. This is likely due in part to broad inclusion criteria, but false negatives may also be occurring because important features such as gliosis, blood products in the resection cavity, and brain shift are not included in simulated data.

Researchers have also explored automated methods for resection zone segmentation that do not rely on machine learning. Casseb et al. developed ResectVol (53), an SPM-based program that relies on differences in tissue probability maps between preoperative and postoperative images to identify the resection site and estimate removed brain structures. The authors achieved promising results, with a median DSC of 0.77 and significant correlation between automated and manual resection volumes ($r = 0.8$, $p < 0.001$). However, conventional image analysis approaches require longer processing times and may be less resilient to contrast changes associated with ancillary pathology, such as gliosis and edema, or different surgical approaches, such as LITT.

Automating quantification of resected tissue will catalyze the progress in at least four fields of epilepsy and neuroimaging research. First, automated resection segmentation can be used to predict brain network reorganization after surgery. Many patients who are initially seizure free after surgery have a seizure relapse in the long term, possibly due to changes in the networks over time (21). Quantifying brain structures remaining after the surgery, by reliably delineating tissues resected, is crucial in predicting such changes and determining patients who are likely to have late recurrences (21,54). Second, modeling proposed surgeries improves prediction of surgical outcomes (24,55). Precise delineation of resected tissue would facilitate retrospectively analyzing factors associated with seizure outcomes after surgery. Third, our clinical application protocols can be applied prospectively to quantify network changes for alternate surgical strategies before carrying out an actual surgery (25,30). This is particularly useful for patients who are likely to have poor surgical outcomes or in patients in which the site of resection is close to the eloquent cortex.
Finally, surgical resection can cause changes in the position and volume of brain structures between pre- and post-operative timepoints (56) as hematomas evolve, resection cavities contract, and cerebrospinal fluid spaces expand. As such, deformable registration would be required to properly align these images for comparison. However, surgical resection also produces significant errors during deformable registration, resulting in erroneous warping of nearby tissue into the resection cavity (47). Our algorithm can be integrated into neuroimaging pipelines to automatically perform cost-function masking of the resection zone, allowing for more accurate processing of postoperative images.

Our study had several limitations, including a use of strict inclusion criteria, poor representation of laser ablations and extra-temporal resections, and a single image rater. To maximize classifier performance, our initial approach was to select a relatively homogeneous patient population. This limits classifier generalizability to TLE patients and restricts available training data to a smaller sample size. In future studies, criteria will be relaxed to include patients with resections outside the temporal lobe, different surgical approaches, and a broader range of clinical imaging sequences. Furthermore, laser ablations and extra-temporal resection patients were poorly represented in our dataset, which resulted in lower segmentation accuracy for these cases. Increasing representation of laser ablations and extra-temporal resections, either as actual or simulated data, could improve classification for these patients. An additional study limitation was that manual segmentations were only available from a single neuroradiologist. Having a single rater prevents the assessment of inter-rater reliability (IRR) in our study; however IRR has been assessed for resection segmentation by other groups and results are fairly consistent across studies (median DSC 0.84-0.86) (51,57).
2.5 Conclusion

In conclusion, we developed a fully automated method for segmenting the resection cavity and quantifying brain regions removed in TLE surgical patients. Our method performance approaches IRR between radiologists while significantly reducing manual input and can be deployed in an easy-to-use GUI. Automated resection cavity segmentation methods have important implications for predictive models of surgical interventions and consistency across multi-center trials. We openly share all code and model weights for our classifier to enable acceleration towards clinical translation and improvement of epilepsy patient care.
Chapter 3 - Remote effects of temporal lobe epilepsy surgery

Epilepsy has long been associated with cortical atrophy and ongoing seizures are known to cause accelerated cortical thinning. However, the impacts of surgical treatments on the brain and their effects on postoperative cortical thinning in remote brain regions remains poorly characterized. In this chapter, we present a method for assessing postoperative cortical thinning and compare cortical thinning rates between temporal lobe epilepsy patients treated with two common surgical approaches. This manuscript is in preparation for submission to Epilepsia Open and is currently available as a preprint on medRxiv (58), with T. Campbell Arnold as first author.

Remote effects of temporal lobe epilepsy surgery: long-term morphological changes after surgical resection

T. Campbell Arnold¹,², Lohith G. Kini¹,², John M. Bernabei¹,², Andrew Y. Revell²,³, Sandhitsu R. Das⁴, Joel M. Stein⁵, Tim H. Lucas²,⁶, Dario J. Englot⁷,⁹, Victoria L. Morgan⁷,⁹, Brian Litt¹,²,⁴, Kathryn A. Davis²,⁴

1. Department of Bioengineering, University of Pennsylvania
2. Center for Neuroengineering and Therapeutics, University of Pennsylvania
3. Department of Neuroscience, University of Pennsylvania
4. Department of Neurology, University of Pennsylvania
5. Department of Radiology, University of Pennsylvania
6. Department of Neurosurgery, University of Pennsylvania
7. Department of Neurological Surgery, Vanderbilt University Medical Center
8. Department of Radiology and Radiological Sciences, Vanderbilt University Medical Center
9. Institute of Imaging Science, Vanderbilt University Medical Center
Abstract

Objective: We present a semi-automated method for quantifying structural changes after epilepsy surgery and compare the remote structural effects of two surgical approaches, anterior temporal lobectomy (ATL) and selective amygdalohippocampectomy (SAH).

Methods: We studied 37 temporal lobe epilepsy (TLE) patients who underwent resective surgery. Patients were treated with either an anterior temporal lobectomy (ATL, N=22) or a selective amygdalohippocampectomy (SAH, N=15). All patients received same-scanner MR imaging preoperatively and postoperatively (5+ months after surgery). The resection zone was segmented using a semi-automated method with manual review by a neuroradiologist. To analyze postoperative changes in remote brain regions, we modified the Advanced Normalization Tools (ANTs) longitudinal cortical pipeline to account for the resection zone. We compared global and regional annualized cortical thinning between surgical treatments and correlated cortical thinning rates with clinical variables.

Results: Across procedures, there was significant cortical thinning in the ipsilateral insula, pericalcarine cortex, and several temporal lobe regions outside the resection zone. Additionally, cortical thinning was seen in the contralateral hippocampus (p = 0.018, FDR corrected) and an increase in postoperative cortical thickness was seen in the supramarginal gyrus (p = 0.037, FDR corrected). Patients treated with ATL exhibited significantly greater annualized cortical thinning compared to patients treated with SAH (ATL: -0.08 ± 0.11 mm, SAH: -0.01 ± 0.02 mm, t =3.02, p = 0.006). There were significant focal differences between the two treatment groups in the ipsilateral insula (p = 0.037, FDR corrected). No significant differences between ATL and SAH were seen in the contralateral hemisphere. A patients annualized cortical thinning rate was found
to significantly correlated with their age at surgery ($R = -0.33, p = 0.050$) and their preoperative cortical thickness ($R = 0.60, p < 0.001$).

**Significance:** We present and share a semi-automated pipeline for quantifying remote longitudinal changes in cortical thickness after neurosurgery. The technique is applicable to a broad array of applications, including surgical planning and mapping neuropsychological function to brain structure. Using this tool, we demonstrate that patients treated with SAH for refractory temporal lobe epilepsy have less postoperative cortical thinning in remote brain regions than those treated with ATL. We share all algorithm code and results to accelerate collaboration and clinical translation of our work.
3.1 Introduction

Epilepsy affects sixty-five million people worldwide, with one third of patients suffering from drug-resistant epilepsy (DRE) (16,59). Individuals suffering from DRE bear greater risk of premature death, injury and worsening quality of life, including psychosocial dysfunction (60). In patients where a focal seizure onset zone can be localized, surgical resection has been established as an effective treatment for DRE (18). The temporal lobe is the most common site of localization, with temporal lobe epilepsy (TLE) accounting for approximately ~65% of DRE. While anterior temporal lobectomy is the most common surgical approach for patients with TLE, in recent years surgical options have expanded and now include more selective resections (61), focal laser ablations (62), and targeted intracranial neural stimulation (63). These less invasive surgical approaches seek to achieve seizure control while preserving cognitive function and reducing surgical comorbidities (64,65). However, comparisons of cognitive outcomes and seizure freedom between traditional resection and selective amygdalohippocampectomy have been mixed, which likely reflects a complex trade-off between preserving functional brain tissue and removing epileptogenic regions (66–69).

Across institutions, numerous surgical procedures are performed for epilepsy management, including anterior temporal lobectomy (ATL), selective amygdalohippocampectomy (SAH), lesionectomy, and focal ablation. Substantial evidence, including a rigorous randomized controlled trial and multi-institution meta-analyses, indicate that ATL is superior to continued antiepileptic medication in up to 70% of cases (18,70). However, evidence indicates that more focal surgery, such as laser ablations, may lead to better cognitive outcomes (62,69). Despite the growing list of surgical options, clinicians lack quantitative methods for assessing the impact of different procedures on brain structure and function. It is known that ongoing seizures are associated with accelerated atrophy throughout the brain, which contributes to cognitive decline (71). However,
comparatively little is known about how surgical procedures affect long-term brain morphometry. Quantitative neuroimaging could provide key insights into differences in long-term seizure-freedom and cognitive outcome between surgical treatments thus providing clinical guidance and improve patient outcomes (72).

Surgical resection has remote consequences outside of local tissues that are removed, such as white matter atrophy (73), however these remote changes are not well understood. While many studies delineate the effects of ongoing seizure on cortical thickness (74), few quantify the change seen between presurgical and postsurgical imaging (27,75). It is important to characterize downstream effects of surgical procedures as they may negatively affect patient outcomes and lead to decreased rates of seizure freedom or increased cognitive side-effects from surgery. Furthermore, given a planned resection by the clinical team, quantitative measurements in combination with brain connectivity analyses could be used to model the remote effects of a proposed surgery on the brain (25,30). Recent neurosurgical advances have refined resective therapies for DRE, but there remains a great need to better understand the effects of surgery on the postsurgical brain to guide future therapy.

In this study, we use advanced computational imaging tools to quantify changes throughout the brain and hypothesize that there are measurable changes in cortical thickness caused by epilepsy surgery. We propose that these changes are a function of the surgical technique employed. To understand the effects of surgical approach on brain structures, we developed quantitative techniques to evaluate postoperative imaging for cortical thickness changes in adjacent and remote brain regions. We quantify downstream effects of two common surgical procedures, ATL and SAH, in 37 patients who underwent temporal lobe epilepsy surgery. We developed a semi-automated method for segmenting resection cavities and considering the extent of the resection zone when computing cortical thinning in the rest of the brain. We publicly share our code to allow others to
validate and improve upon our methods for assessing post-operative changes in cortical thickness to accelerate translation into clinical practice.

3.2 Methods

3.2.1 Patients

Thirty-seven patients who underwent surgery for localization-related epilepsy across two institutions, Hospital of the University of Pennsylvania (HUP, N=14) and Vanderbilt University Medical Center (VUMC, N=23), were recruited for this study. The protocol was approved by the Vanderbilt University Institutional Review Board and the Institutional Review Board of the University of Pennsylvania. All participants gave informed consent. Patients were age-matched across institution (HUP: 38.9 ± 13.3, VUMC: 39.5 ± 12.7) and were treated with either an ATL (N=22) or SAH (N=15). Clinical variables collected include gender, age at surgery, age at onset, disease duration, seizure lateralization, years since surgery, and surgical outcomes. For patients treated at HUP, age at onset, disease duration, and surgical outcomes were not available, and all related statistics only include patients from VUMC. For surgical outcomes, patients were assessed using the Engel classification system and binarized into Engel I (more favorable outcome) and Engel II-IV (less favorable outcome) groups. One patient treated with a focal laser ablation of the amygdala and hippocampus was included in the SAH group. Patient demographics and clinical characteristics are listed in Table 3.
<table>
<thead>
<tr>
<th>Procedure</th>
<th>ATL</th>
<th>SAH</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of participants</td>
<td>22</td>
<td>15</td>
<td>37</td>
</tr>
<tr>
<td>Gender (Female / Male)</td>
<td>14/8</td>
<td>8/7</td>
<td>22/15</td>
</tr>
<tr>
<td>Age at surgery (Mean ± SD)</td>
<td>39.6 ± 12.6</td>
<td>34.8 ± 12.6</td>
<td>37.7 ± 12.0</td>
</tr>
<tr>
<td>Age at onset (Mean ± SD)</td>
<td>25.6 ± 10.1</td>
<td>15.3 ± 11.3</td>
<td>18.9 ± 11.9</td>
</tr>
<tr>
<td>Duration (Mean ± SD)</td>
<td>17.4 ± 13.8</td>
<td>24.2 ± 15.7</td>
<td>21.8 ± 15.4</td>
</tr>
<tr>
<td>Seizure side (Left / Right)</td>
<td>7/15</td>
<td>6/9</td>
<td>13/24</td>
</tr>
<tr>
<td>Years since surgery (Mean ± SD)</td>
<td>1.7 ± 1.3</td>
<td>2.8 ± 1.3</td>
<td>2.1 ± 1.4</td>
</tr>
<tr>
<td>1 Year outcome (Engel I / Engel II-IV)</td>
<td>5/3</td>
<td>11/4</td>
<td>16/7</td>
</tr>
<tr>
<td>Latest outcome (Engel I / Engel II-IV)</td>
<td>5/3</td>
<td>10/5</td>
<td>15/8</td>
</tr>
</tbody>
</table>

**Table 3 - Demographics of patients included in study.** Notes: One patient was treated with a laser ablation of left amygdala & hippocampus and was included in the SAH group for this study. Some clinical variables, including age at onset, disease duration, and surgical outcome are only available for patients treated at VUMC (N=23). Abbreviations: ATL, Anterior Temporal Lobectomy; SAH, Selective Amygdalohippocampectomy; SD, Standard Deviation.

### 3.2.2 Image Acquisition

All patients underwent a clinical epilepsy neuroimaging protocol both preoperatively and postoperatively (>5 months after surgery) as part of their standard care. The post-resection imaging protocol was acquired, on average, 26 months after implant and resection. For each patient, 1 mm isotropic pre-implant T1-weighted MRI (T1w) and post-resection T1w images were acquired. Preoperative and postoperative images were collected on the same scanner to avoid the confounds of inter-scanner variability. Patients at HUP were imaged using a Siemens Trio 3T scanner with a 64-channel head coil and received a T1w scan with the following acquisition parameters (1 mm isotropic, TE = 3.87ms, TR = 1.62s, flip-angle = 15). Patients at VUMC were imaged using a Phillips 3T scanner with a 32-channel head coil and received a T1w scan acquired with the following acquisition parameters (1 mm isotropic, TE = 4.61ms, TR = 8.9ms, flip-angle = 8).
3.2.3 Selection of surgical approach

At VUMC, patients are typically chosen for SAH when there is a high level of confidence that the seizures are mesial (e.g. seemingly mesial semiology, MTS on MRI, hippocampal onset if the patient undergoes intracranial implant). ATL is typically performed when clinicians are relatively confident that the seizures are localized to the unilateral temporal lobe, but there is less evidence and confidence that they are mesial (e.g. atypical semiology, nonlesional MRI, possible lateral involvement if the patient undergoes intracranial implant). All HUP patients included in this study were treated with ATL.

3.2.4 Resection Segmentation

Patient images were analyzed using Advanced Normalization Tools (ANTs) (32,45). Preoperative images were registered to the postoperative image and the Atropos algorithm (76) was used to segment images into six tissue types: cerebral spinal fluid (CSF), grey matter (GM), white matter (WM), deep grey matter (DGM), cerebellum, and brainstem. In postoperative images, the resection area is predominantly labeled as CSF, while the same voxels would be labeled as GM or WM in preoperative images.

To generate an initial, automated estimation of the resection zone, we subtracted postoperative and preoperative CSF probability maps. Prior to subtraction, a gaussian smoothing kernel (sigma = 2) was passed over the probability maps to avoid large contrast due to misalignment (77). Voxels with greater than 25% increases in CSF probability from preoperative to postoperative timepoints were considered as candidates for the resection mask. We constrained the segmentations to only include voxels original classified as GM or WM in preoperative images (78). To eliminate stray voxels, only the largest contiguous cluster of voxels in the resection mask was retained. All
resection zones estimations were manually reviewed and corrected by a neuroradiologist (JMS, 8 years of experience) using ITK-SNAP (35).

To compare the extent of resection associated with each procedure, probability maps were generated for each combination of surgical technique (ATL or SAH) and hemisphere (left or right). For each subject, preoperative and postoperative T1w images were rigidly coregistered and resection masks were transformed into preoperative space. Preoperative imaging was then registered to the MNI ICBM152 template (79) and the transformation was applied to the resection mask, warping it into MNI space. A probability map with values ranging from 0% to 100% was generated by averaging resection masks in MNI space across all patients within each procedure and hemispheric combination.

3.2.5 Cortical Thinning Estimation

We assessed cortical thickness using an adaptation of the well validated ANTs longitudinal cortical thickness pipeline (32, 45, 80–83). For patients with cortical resections, the standard ANTs cortical thickness pipeline causes significant image distortion during deformable registration of postoperative and template images (84, 85). This distortion results from white matter and gray matter in the postoperative image erroneously being pulled into the resection cavity to better match the template image intensity values (Figure 7), and ultimately leads to downstream inaccuracies in cortical thickness estimation.
Recently, a new deep learning enhanced version of ANTs was released (ANTsPyNet), which eliminates the need for deformable registration during cortical thickness estimation (32). We further modified the longitudinal cortical thickness pipeline from this package to account for the postoperative resection zone during single-subject template (SST) generation. When generating an SST, voxel intensity values are reduced in the resection zone because preoperative and postoperative images are averaged. We used the resection segmentation to impute preoperative values into the affected region of the SST. This generates a template image that closely aligns with the patients imaging, without reduced intensity values in the resection zone.

Mean cortical thickness was estimated for preoperative and postoperative images across a set of 62 regions in the Desikan-Killiany-Tourville (DKT) atlas (46,86). The ipsilateral amygdala, hippocampus, entorhinal cortex, and parahippocampal regions were excluded from our analysis, as these regions are completely or nearly completely removed during resective surgery. Cortical thinning was calculated as the difference of mean postoperative and preoperative cortical thickness. To account for differences in followup time between subjects, we divided by the interscan interval to yield annualized cortical thinning rates. A visual description of the analytical
pipeline is provided in Figure 8. Patterns of cortical thinning were visualized using the BrainNet Viewer toolbox (87).

Figure 8 - Analytical pipeline. Pre- and post-operative imaging for 22 patients treated with ATL and 15 patients treated with SAH were processed through a modified version of the ANTs cortical thickness pipeline. Subsequently each brain region is compared in group level statistical tests. Abbreviations: ATL, Anterior Temporal Lobectomy; SAH, Selective Amygdalohippocampectomy.

3.2.6 Statistical Analysis

The two treatment groups, ATL and SAH, were compared for differences in age at surgery, gender, interscan interval, and lateralization using a two-sample t-tests or chi-squared tests. Global differences in cortical thinning were compared between groups using a two-sample t-test. Differences in cortical thinning were assessed across 62 individual brain regions from the DKT
atlas. Brain regions were assessed as either ipsilateral or contralateral to the resection site. Across brain region comparisons, false-discovery rate (FDR) correction was applied to account for multiple comparisons. Pearson’s correlation and t-tests were used to relate cortical thinning to clinical variables.

3.2.7 Code Availability

All code related to resection segmentation, cortical thickness estimation, and statistical analysis are available at: https://github.com/tcama/remote_effects

3.3 Results

3.3.1 Demographics

Thirty-seven patients were included in this study (14 patients from the Hospital of the University of Pennsylvania and 23 patients from the Vanderbilt University Medical Center). One patient was excluded due to motion degradation in postoperative imaging. Table 1 summarizes important clinical demographics for the study patients. For patients treated at HUP, age at onset, disease duration, and surgical outcome were not available. There was no statistically significant difference between patients treated with SAH or ATL in patient age at surgery (two-sample t-test: $t = 1.13, p = 0.27$), gender ($\chi^2 = 0.39, p = 0.82$), or side of seizure onset ($\chi^2 = 0.26, p = 0.87$). There was a significant difference in the interscan interval, or time between surgery and postoperative imaging (two-sample t-test: $t = 2.64, p = 0.012$), such that patients treated with ATL had a shorter interval ($1.7 \pm 1.3$ years) than those treated with SAH ($2.8 \pm 1.3$ years).
3.3.2 Preoperative cortical thinning

To assess for selection bias between scan sites and surgical treatment groups, we compared mean preoperative cortical thickness for patients treated with SAH at VUMC, ATL at VUMC, and ATL at HUP. For patients at VUMC, there was no significant difference in preoperative cortical thickness for those treated with SAH or ATL (two-sample t-test: t = 1.82, p = 0.083), indicating both treatment groups had similar preoperative states (Figure 9A). However, HUP patients treated with ATL had significantly lower mean cortical thickness compared to VUMC patients treated with SAH (two-sample t-test: t = 4.46, p < 0.001) and ATL (two-sample t-test: t = 3.50, p = 0.002). Regional patterns of cortical thickness appear to be similar between all groups, although with a global decrease in thickness across all regions in HUP patients (Figure 9B-D). Given that there was no significant difference in age between any combination of scan sites or treatment groups, it is possible the lower cortical thickness values may be attributable to intra-scanner differences. Additionally, cortical thinning is expected to be greater in patients with longer interscan intervals, which could bias analyses towards greater cortical thinning in the SAH treatment group. To address the discrepancy in interscan interval and between scanner sites, all subsequent analyses will compare annualized cortical thinning rates.
Figure 9 - Preoperative cortical thickness. (A) Patients treated at HUP with ATL had significantly lower preoperative cortical thickness compared to SAH (two-sample t-test: t = 4.46, p < 0.001) and ATL (two-sample t-test: t = 3.50, p = 0.002) patient treated at VUMC. (B-D) Regional preoperative cortical thickness values can be seen patients treated at (B) VUMC with SAH, (C) VUMC with ATL, and (D) HUP with ATL. Cortical thickness patterns are similar for both VUMC groups, while patients treated at HUP with ATL have markedly lower cortical thickness values globally. Abbreviations: ATL, Anterior Temporal Lobectomy; SAH, Selective Amygdalohippocampectomy; HUP, Hospital of the University of Pennsylvania; VUMC, Vanderbilt University Medical Center.
3.3.3 Resection Extent

To illustrate the extent of resections, we generated probability maps for each combination of surgical technique and hemispheric lateralization, as seen in Figure 10. Surgical margins are based on the patient’s language lateralization, with wider margins for epilepsy surgery in the non-dominant hemisphere (88). The left hemisphere is typically the language dominant hemisphere, which results in larger resection margins being associated with right hemisphere surgeries in our dataset. In both hemispheres, patients treated with SAH have reduced resection margins compared to patients treated with ATL.

**Figure 10 - Probability maps for each surgical group.** For each surgical technique (SAH or ATL) and hemisphere combination (left or right), a probability map was generated to visualize the extent of the resection procedures. Each subject’s T1w imaging was registered to the MNI ICBM152 template and a manually drawn resection mask were warped to MNI space. A probability map with values ranging from 0% to 100% was generated by averaging the resection masks in MNI space. Warmer colors indicate areas where more patients treated with a surgical procedure had surgical resection overlap. In both hemispheres, patients treated with ATL have the highest overlap in the anterior temporal pole, with probability of resection reducing as you travel posteriorly. Patients treated with SAH have the largest overlap in the mesial temporal lobe and have a reduced resection coverage compared to ATL patients.
3.3.4 Cortical Thinning

In the hemisphere ipsilateral to surgical resection, across both procedures significant cortical thinning was seen in several temporal lobe regions adjacent to the surgical site (Figure 11A), including the inferior temporal gyrus (p = 0.008, FDR corrected), middle temporal gyrus (p = 0.034, FDR corrected), superior temporal gyrus (p = 0.027, FDR corrected), and transverse temporal gyrus (p = 0.037, FDR corrected). Ipsilateral cortical thinning was also seen in the fusiform cortex (p < 0.001, FDR corrected), pericalcarine (p = 0.034, FDR corrected) and insula (p = 0.003, FDR corrected). Interestingly, there was also significant cortical thinning in the contralateral hippocampus (p = 0.015, Figure 11B).

Figure 11 - Changes in postoperative cortical thickness. This figure illustrates significant cortical thinning findings. (A) Ipsilateral cortical thinning across both procedures was seen in temporal lobe regions, the insula, the fusiform cortex, and the pericalcarine cortex. A cortical thickness increase was seen in the supramarginal gyrus. (B) In the contralateral hemisphere, cortical thinning was seen in the hippocampus. (C) Globally there was significantly greater cortical thinning in patients treated with ATL, (D) with a focal difference in the insula.

To assess whether there was a global difference in cortical thinning between patients treated with ATL and SAH, annualized cortical thinning was compared between groups using a
two-sample t-test (Figure 11C). Globally, there was significantly more cortical thinning in patients treated with ATL over those treated with SAH (two-sample t-test: $t = 3.02, p = 0.006$). Patients treated with ATL had elevated annualized cortical thinning ($0.08 \pm 0.11$ mm per year), while those treated with SAH had much lower rates of cortical thinning ($0.01 \pm 0.02$ mm per year). There was no global difference in cortical thinning based on left or right hemispheric lateralization.

At the individual brain region level, there was significantly more cortical thinning in the ipsilateral insula for patients treated with ATL ($p = 0.037$, Figure 11D). There were no differences between ATL and SAH in the contralateral hemisphere. Although statistical differences in many regions did not persist after FDR correction, overall the extent of cortical thinning was greater in patients treated with ATL compared to SAH, as visualized in Figure 12, with the notable exception of the supramarginal gyrus.

**Figure 12 - Differences in cortical thinning rates between ATL and SAH.** This figure compares cortical thinning rates between patients treated with ATL and SAH at the level of individual brain regions. Areas with cool colors indicate greater cortical thinning in patients treated with ATL, while warmer colors indicate areas with greater cortical thinning in SAH patients. Globally, rates of cortical thinning are higher for patients treated with ATL, with the largest differences in the
ipsilateral insula, ipsilateral temporal lobe outside the resection zone, and contralateral hippocampus. The supramarginal gyrus is a notable exception, where rates appear to favor ATL.

### 3.3.5 Relating cortical thinning to clinical variables

To better understand the factors that contribute to accelerated cortical thinning, we correlated individual thinning rates with clinical variables of the patients. We analyzed their age at surgery, age at onset, disease duration, and mean preoperative cortical thickness. We found that an individual’s annualized cortical thinning rate was associated with their age at surgery ($R = -0.33$, $p = 0.050$) and their preoperative cortical thickness ($R = 0.60$, $p < 0.001$) (Figure 13). There was also a trending association with disease duration ($R = -0.40$, $p = 0.062$), though no relationship with age at onset ($R = 0.17$, $p = 0.42$). We also compared cortical thinning rates between patients that had more favorable outcomes (Engel I) and less favorable outcomes (Engel II-IV). There was no significant difference between surgical outcome measured 1 year after surgery ($p = 0.30$) or the most recent Engel outcome available ($p = 0.26$).

**Figure 13 - Correlations between clinical variables and cortical thinning.** This figure illustrates correlations between cortical thinning and a patient’s age at surgery ($R = -0.33$, $p = 0.050$) and their preoperative cortical thickness ($R = 0.60$, $p < 0.001$).


3.4 Discussion

Epilepsy has long been associated with brain atrophy, particularly hippocampal sclerosis in patients with temporal lobe epilepsy (89,90). Over the past two decades, it has become well established that cortical atrophy in epilepsy is not confined to the epileptogenic zone but rather can be observed throughout the neocortex (91–95). While the effects of ongoing seizures on cortical thickness have been relatively well studied, comparatively little is known about how surgical operations to treat epilepsy subsequently affect cortical thinning (27,75). There is a lack of focus on postsurgical treatment groups in part because of the difficulties in working with postoperative imaging, whereby extra precautions must be taken to ensure that structural alterations do not adversely affect analyses. In this study, we conducted a whole-brain comparative analysis of longitudinal cortical thinning between two epilepsy surgical techniques, anterior temporal lobectomy and selective amygdalohippocampectomy. To accomplish this goal, we developed semi-automated software for assessing longitudinal cortical thinning while accounting for the tissue resected during surgery, which we publicly release.

Our most important finding was that patients treated with SAH had significantly less postoperative cortical thinning than those treated with ATL. While there was a significant effect globally, further analysis revealed that the largest driver was located in the ipsilateral insula. Additionally, across both procedures we found significant cortical thinning in ipsilateral temporal lobe regions outside the resection zone, as well as in the ipsilateral fusiform cortex, insula, and pericalcarine cortex. Cortical thinning was also seen in the contralateral hippocampus. There was also one region, the supramarginal gyrus, that showed a postoperative cortical thickness increase. We found that age-related variables, such as age at surgery and preoperative cortical thickness, were predictive of postoperative cortical thinning rates.
We speculate that cortical thinning patterns and regional postoperative differences based on surgical approach are reflective of the underlying network connectivity between the affected cortex and downstream regions (96). Temporal and limbic brain regions are known to be integrated into a functional brain network and in prior studies TLE cortical thinning patterns have been subjectively observed to overlap with regions structurally connected to the ipsilateral hippocampus (97,98). Recent work has revealed that epilepsy accelerates cortical atrophy and that regions connected to the epileptogenic zone are most likely to be afflicted with cortical atrophy (97,99). The overlap between cortical thinning patterns and structural connectivity maps of the putative epileptogenic zone raises the possibility of using cortical thinning as a method for localizing the epileptic focus or differentiating epilepsy subtypes (100).

Our study adds to a growing body of work that validates cortical thickness measurements as an important biomarker for epilepsy. In a recent longitudinal case-control study, Galovic et al. observed that epilepsy patients had a two-fold increase in the rate of annualized cortical thinning when compared to healthy age-matched controls (TLE 0.024 ± 0.061 mm per year vs healthy controls 0.011 ± 0.029 mm per year) (97). In our own work, we observed that patients treated with ATL had even higher rates of annualized cortical thinning 0.08 ± 0.11 mm per year, while those treated with SAH showed minimal progressive cortical thinning (0.01 ± 0.02 mm per year) similar to the rate of cortical thinning seen in healthy controls in the Galovic et al. study. Patients treated with ATL demonstrated annualized cortical thinning that was 3 times greater than DRE patients from the Galovic et al. study. The accelerated cortical thinning and differences between the surgical approaches may result from greater Wallerian degeneration and loss of functional connections in patients treated with ATL, though tractography studies would be necessary to confirm this hypothesis.

One exciting potential for tools that assess postoperative cortical thinning is the possibility of relating quantitative structural analyses and cognitive outcomes. Several studies have compared
selective and traditional surgical approaches, with most citing similar seizure outcomes and superior cognitive outcomes for selective procedures (64,65). However, the cause of these differences in cognitive outcomes remains poorly understood. Selective procedures may preserve more functional gray matter near the focal lesion, or perhaps traditional resections cause greater neurologic injury through trauma to white matter tracts or vasculature, resulting in more widespread effects. In the present study, we demonstrated significant differences in downstream cortical thinning between SAH and ATL. Interestingly, in mesial temporal lobe epilepsy patients Bettus et al. observed a decrease in ipsilateral temporal lobe connectivity while conversely connectivity increased in the contralateral hemisphere (101). They proposed this may reflect a compensatory mechanism on the contralateral side to preserve normal cognitive function. Furthermore, recent work has demonstrated long-term functional connectivity changes in the contralateral hippocampus after surgery (102). In our own data, we saw cortical thinning in the contralateral hippocampus, which could be a structural analog to functional observations from other studies. However, the tools outlined in this work would need to be paired with postoperative neuropsychiatric outcomes and functional or structural connectivity to better understand the relationship between postoperative cortical thickness and cognitive outcomes.

Our quantitative cortical thickness methods could provide clinical utility during the surgical planning process, both for epilepsy surgery and other applications. Network-based models provide an exciting in silico approach to mapping seizure onset zones and resection zone targeting. New methods, like virtual resection, allow clinicians to map the role of each node using network-control methodology (29,30). In this case, nodes are representations of electrodes as part of the ECoG or sEEG implant. However, these methods focus solely on the tissues being resected and do not account for downstream effects on the network. The brain regions identified here could be used in combination with a normative approach to modify network-based models to account for such downstream effects (103). Additionally, recent work by da Silva et al. has demonstrated that
epilepsy surgery is associated with changes to structural connectivity in regions remote to the surgical site (104). By using lesional models of structural connectivity in combination with patterns of postoperative cortical thinning, we can increase our understanding of how focal treatments affect brain structure globally.

Three major limitations of this study are the cohort size, lack of clinical variables, and heterogeneity in surgical approach and interscan interval. These factors are often limitations in studies of drug resistant epilepsy as obtaining consistent longitudinal data can prove difficult and patients are often lost to follow-up. The present study may be underpowered to detect subtle differences in cortical thinning, especially in subjects with shorter interscan interval times. While the study may be underpowered, we have still applied a statistical correction in an attempt to minimize the potential of Type I error. Additionally, annualized cortical thinning analyses were used to address heterogeneity in the interscan interval and differences in baseline cortical thickness. This study represents an initial foray to uncover possible regions of interest that clinicians should note as likely to be affected by resection, however larger studies with consistent scan times and recruitment practices should be conducted to establish definitive differences between the treatment groups. Additionally, a comparison to age-matched controls would be desirable, as this would help establish whether patients treated with SAH return to cortical thinning rates similar to the healthy population, which cannot be assessed in the present study. Finally, an additional limitation is the lack of behavioral outcome measures, such as post–resection neuropsychological variables (e.g. memory and language). Future studies should include these outcome variables in addition to new surgical methods such as laser interstitial thermal therapy (LITT), now that our algorithms and methods have been developed and openly shared.
3.5 Conclusion

We provide evidence for a relative reduction in postoperative cortical thinning for patients treated with more focused surgical resections. Cortical thinning was globally increased for patients treated with resections, with a significant focus in the ipsilateral insula. Anticipated cortical thinning can inform surgical decisions and postoperative cognitive rehabilitation. In addition, we provide an open-access semi-automated pipeline for the assessment of longitudinal changes in cortical thinning for patients with large structural abnormalities, such as brain tumors or surgical resections.
Recent advances in hardware and software have led to the development and commercial deployment of new low-field MRI devices. These devices offer significant advantages, including cost savings, portability, and reduce siting and safety considerations. However, low-field MRI also faces unique challenges as these devices are deployed clinically. In this chapter, we review potential clinical applications of low-field MRI and identify the opportunities and challenges facing the technology. This manuscript was accepted and is in press at the *Journal of Magnetic Resonance Imaging*, with T. Campbell Arnold as first author.

**Low-field MRI: Clinical promise and challenges**

T. Campbell Arnold¹², Colbey W. Freeman³, Brian Litt²⁴, Joel M. Stein²³

1. Department of Bioengineering, University of Pennsylvania
2. Center for Neuroengineering and Therapeutics, University of Pennsylvania
3. Department of Radiology, University of Pennsylvania
4. Department of Neurology, University of Pennsylvania
4.1 Introduction

Magnetic resonance imaging (MRI) is a mainstay of modern medicine and has led to significant advances in basic science, clinical diagnostics, and patient care. As such, MRI is widely utilized in high income countries (HICs), with 1.9 MRI scans annually per every 10 American Medicare enrollees (105). Because of the high cost and technical barriers associated with MRI, adoption in low-and-middle income countries (LMICs) has been challenging (106). Approximately ninety percent of people do not have access to MRI (9), and an astounding two-thirds of the world’s population does not have access to even basic medical imaging (10–12). Even in HICs such as the United States, MRI is unavailable in many rural areas or to patients with significant disability or device constraints (107,108). Recent advances in low-field MRI offer potential solutions, with the promise of cheaper MRI devices that can be deployed in novel settings. However, low-field MRI comes with its own unique challenges, and it remains to be seen how these devices will be deployed clinically. In this review, we take a critical view of clinical applications of low-field MRI and identify potential promise and relevant challenges facing the technology.

The article is predominantly organized by clinical application. We begin with a brief overview of recent hardware and software advances, which have sparked renewed interest in clinical low-field devices. Next, we discuss work in five clinical domains where low-field MRI offers significant clinical promise: acute and ICU neuroimaging, outpatient neuroimaging, MRI-guided procedures, pediatric imaging, and musculoskeletal imaging. In each section we provide a contextual overview of the literature and discuss current standards of care. We then review innovations and promising research trends in low-field imaging. Finally, we discuss the possible application and challenges facing low-field imaging. Although this review only covers a fraction of possible use cases, our goal is to convey the potential clinical impact of the technology and stimulate further clinical translation.
4.2 Hardware & Software advances

The definition of “low-field” varies in the literature, sometimes broadly referring to anything below 1.5T while at other times indicating a narrow magnetic field strength band between 0.01T-0.1T. For the purposes of this review, we will refer to devices using the following distinctions (Figure 14): Ultra-low-field (ULF) ≤ 0.01T < very-low-field (VLF) < 0.1T ≤ low-field (LF) ≤ 0.3T < mid-field (MF) ≤ 1.0T < high-field (HF) ≤ 3T < very-high-field (VHF) < 7T ≤ ultra-high-field (UHF). We will use “lower-field” to refer to devices below standard 1.5T clinical scanners and “higher-field” for 1.5T and above. As researchers explore different fields strengths, new levels of distinction will be developed to better communicate; however, we can obviate confusion by defining terminology in our work or developing standards through professional societies, such as the International Society of Magnetic Resonance Medicine (ISMRM).

Figure 14 - Defining “low-field” MRI. This figure defines how scanners at difference field strengths will be categorized in this article, with the following boundaries illustrated on the bottom: Ultra-low-field (ULF) ≤ 0.01T < very-low-field (VLF) < 0.1T ≤ low-field (LF) ≤ 0.3T < mid-field (MF) ≤ 1.0T < high-field (HF) ≤ 3T < very-high-field (VHF) < 7T ≤ ultra-high-field (UHF). Select commercially available scanners with a field strength of 1 tesla or lower are illustrated on the top.
Recent renewed interest in lower-field strength devices has been sparked by hardware improvements, including improved gradient, coil, and magnet designs, as well as software developments, such as noise reduction and improved image reconstruction. Additionally, commercial interest from MR vendors has led to FDA clearance of several lower-field systems since 2018, including the 1T Aspect Embrace neonatal scanner, 0.064T Hyperfine Swoop head scanner, 0.066T Promaxo prostate scanner, 0.55T Siemens Magnetom Free.Max general purpose scanner, and 0.5T Synaptive Evry intraoperative scanner. However, it is important to recognize that low-field MRI is not a new development. In fact, during MRI’s infancy in the 1980s, papers using the terms low-field MRI and high-field MRI were cited with equivalent frequency. A gap emerged after 1985, when the first 1.5T scanners were introduced. This citation gap grew throughout the 1990s, but began to widen significantly in the early 2000s, when 1.5T scanners were becoming the clinical standard. Higher signal-to-noise ratio (SNR) permitting faster imaging, higher resolution, greater contrast sensitivity, and more advanced sequences helped higher-field scanners gain a dominant market share. Low-field scanners have remained available and in use through this time, however they have largely been relegated to niche use cases.
Figure 15 - Research interest in low-field. The relative number of PubMed citations (109) for high-field MRI (blue) and low-field MRI (red) have been diverging in recent decades, reflecting the dominance of high-field scanners.

While higher-field devices won market dominance based on an argument of higher image quality, the recent lower-field strength renaissance has been predominantly driven by another factor: cost (112). Medical care in the United States has risen dramatically in recent years, with medical imaging being a significant contributing factor. High-field MRI devices are expensive and their cost has increased over time. This increasing cost is burdening the US health care system and also makes MRI out of reach for much of the world’s population. Increased healthcare costs in the United States and lack of medical imaging access around the world have led to renewed interest in cost-effective MRI designs. The largest component of MRI device cost is the magnet, with total device cost being roughly 1 million US dollars per tesla of magnetic field strength (113). Lower-field strength devices could offer significant device cost savings.

Although lower-field MRI is associated with lower image quality, it is actually image SNR per unit time that is proportionate to magnetic field strength. Stronger magnets reduce the time necessary to achieve a certain level of sensitivity (112). Given sufficient time, lower magnetic fields can produce high-SNR images of diagnostic quality; however, MRI acquisitions must generally
fall within a reasonable scan time to be tolerated by patients and provide efficient clinical care. Recent software and hardware advances have dramatically improved the quality of images per unit time, making it easier to image at lower-field strengths within clinically feasible timeframes.

Commercial and clinical translation of low-field technology has been enabled by the recent proliferation of low-field devices from numerous research groups using various scanner designs, including the ultra-low-field 6.5mT electromagnet scanner at MGH (113), the 80 mT and 50 mT very-low-field Halbach array devices respectively at MGH and the University of Leiden (114,115), a fast field cycling scanner at the University of Aberdeen that can operate at any magnetic field strength between 50 micro-T and 200 mT (116), and the 55mT very-low-field permanent dipole system at the University of Hong Kong (117). A common theme is MRI design simplification to facilitate scanner production, maintenance, and operation in low-resource settings (118). Additionally, reduced weight and siting requirements for some low-field devices has enabled them to be portable.

At lower-field strengths, coil noise is dominant, necessitating the development of low-cost, 3D-printed head coils with optimized wire diameter, spacing, and windings to produce homogenous magnetic fields (113). For portable lower-field systems, radiofrequency (RF) shielding methods must be lightweight. Liu et al. and Su et al. used passive coils to predict and remove electromagnetic noise (117,119). Another trend in lower-field devices is reduction in reliance on gradient coils, which require high amounts of power. Cooley et al. designed a cylindrical Halbach array, a scanner design composed of multiple small permanent magnets, with optimized magnet placement resulting in a built-in readout field gradient with minimal stray flux outside the bore (114). Importantly, this removes the need for one of the gradient systems and lowers the power and cooling requirements of the system. Additionally, the group has leveraged a rotating scanner bore to collect 2D images without the use of any gradient coils, thereby permitting silent imaging (120). Another approach
has been to step-down high-field systems, operating these systems at lower-field strengths while maintaining other state-of-the-art commercial features, such as gradients and coils (121,122). Researchers can then explore lower-field imaging capabilities for research purposes without redesigning the entire scanner. Campbell-Washburn et al. has taken this approach at the NIH, stepping a 1.5T scanner down to 0.55T. This approach permits high-quality imaging with expanded opportunities in procedural device compatibility and has led to the commercial development of the 0.55T Siemens Free.Max system.

On the software side, there have been steady improvements in image acquisition, reconstruction, and post-processing. These software advances have been facilitated by advancements in deep learning, increased graphics processing unit (GPU) availability, and the open-source movement. At low-field strengths, there is a trade-off between SNR, resolution, and scan time. To maintain clinically relevant scan times, researchers have sought rapid imaging methods, such as subsampling in the sensor space, but this results in noise and image artifacts after conventional reconstruction into the image domain. Recently, compressed sensing and deep learning have enabled reconstruction of MR images from a small subset of k-space while maintaining high image quality, permitting faster acquisitions with high fidelity (123–125). Postprocessing of data in the imaging domain has greatly benefited from deep learning, including applications in super-resolution (126–128), segmentation (129), denoising (130), and artifact rejection (131). The low-field research community has also engaged with the open source movement, most notably through the Open Source Imaging Initiative (OSI: opensourceimaging.org) (132,133). Projects include both hardware, including the 50mT Halbach array magnet (115,134), and software designs, such as the MaRCos acquisition control system (135).
While it is crucial to acknowledge the recent hardware and software advances that enable clinical translation, a full discussion of the topic is beyond the scope of this review. For more information on advances in low-field MRI technology, the authors recommend the excellent reviews by Wald et al. & Marques et al. (9,136). The remainder of this review focuses on potential clinical applications of low-field MRI devices.

4.3 Acute & ICU Neuroimaging

In critical care patients, neuroimaging can identify acute problems such as stroke and traumatic or nontraumatic hemorrhage which require immediate intervention. In this patient population, clinicians are monitoring for large changes, changes over time, and actionable acute findings. Importantly, these patients often cannot be transported easily outside the intensive care unit (ICU). Stroke is a leading cause of morbidity and mortality worldwide, causing a loss of 110 million disability-adjusted life-years and an estimated 6.5 million deaths each year (137,138). The two main types of strokes are ischemic and hemorrhagic, with ischemic strokes accounting for 87% of cases in the United States (139,140). Differentiating between ischemic and hemorrhagic strokes is an essential first step towards treatment. Early evaluation of stroke is essential, as the irreversible core of the infarct enlarges with time, and evidence supports a 3-4.5-hour treatment window for intravenous thrombolysis and up to a 24-hour window for mechanical thrombectomy, after which outcomes are considerably worse (141,142).

Neuroimaging is the dominant method for determining stroke subtype, with computed tomography (CT) and MRI being the modalities of choice (143). Provided that MRI is readily accessible, the American Academy of Neurology recommends MRI over CT because it avoids ionizing radiation and has superior soft-tissue contrast facilitating detection of smaller infarcts.
Diffusion-weighted imaging (DWI) has high sensitivity and specificity for detection of ischemia, with other sequences, such as fluid-attenuated inversion recovery (FLAIR) and gradient echo (GRE), providing complementary information. While MRI has been shown to be diagnostically superior to CT, conventional MRI can be significantly more expensive, is not always readily available, and is contraindicated in ~10% of patients.

Low-field MRI may address many of the challenges preventing wider adoption of MRI for stroke imaging. First, low-field scanners are often less expensive to purchase and operate than conventional MRI. Commercial systems are available for ~75K/year USD, with several research groups reporting the development of systems under 20K USD. The lower cost combined with the portability of lower-field systems has the potential to increase availability. Finally, many patients with conventional MRI contraindications, such as pacemakers, defibrillators, intravenous medication pumps, deep brain stimulators, vagus nerve stimulators, and cochlear implants, can be safely imaged on some lower-field systems. Given that stroke patients may be incapacitated during imaging, the reduced screening requirements could be a substantial benefit.

In the 1990s and early 2000s, several studies explored the development of diffusion-weighted and perfusion-weighted sequences for stroke diagnosis on lower-field scanners (0.1-1.0T range). These studies employed fixed MRI systems with either permanent magnets (typically <0.35T) and an open bore or superconducting magnets (typically >0.5T) and a closed bore. While these systems were able to detect many strokes, sensitivity was reduced compared to 1.5T systems or the scan times were not clinically feasible. More modern lower-field MRI machines are equipped with state-of-the-art gradient systems, which could significantly improve DWI in closed-bore systems and make them more useful for stroke imaging.
In very-low-field strength systems (<0.1T), there have been several clinical developments related to stroke imaging in recent years. The first report of stroke imaging on a portable very-low-field MRI was described by Sheth et al. in 2020 (156). The authors used a 0.064T system to image 30 ICU patients with known intracranial abnormalities, including ischemic stroke, hemorrhagic stroke, subarachnoid hemorrhage, traumatic brain injury, and brain tumor. Imaging was performed at the bedside with medical equipment being actively used, including ventilators and dialysis machines. Of the 30 patients, intracranial abnormalities were detected on very-low-field MRI in 97% (29/30, with 1 diffuse subarachnoid hemorrhage missed). In a follow up study, the researchers evaluated 132 participants with intracerebral hemorrhage (ICH, N=50), acute ischemic stroke (AIS, N = 44), or healthy controls (N=38) (157). ICH classification accuracy was 90% (130/144), with a sensitivity of 80% (45/56). AIS and healthy controls were correctly identified as blood negative in 97% of cases (85/88). The volume of manual hematoma segmentations on the 0.064T system strongly correlated with volumes on standard of care (SOC) imaging. Hematoma volume was also found to correlate with cognitive status (rho=0.75/0.8, p<0.001) and functional outcome at discharge (rho=0.59/0.64, p<0.001). An example of a hemorrhage in a 27-year old hydrocephalus patient imaged at our center can be seen in Figure 16A.
The 0.064T system received FDA clearance early in the first months of the COVID-19 pandemic (158). In their initial publication, Sheth et al. imaged an additional 20 ICU patients diagnosed with COVID-19 who presented with altered mental status (156). In this cohort, abnormal findings were present in 40% (8/20) of patients. Another early report on the same system by Turpin et al. described the use of portable MRI in 19 ICU patients, with abnormal findings present in 63% (12/19) of patients. (159) The researchers also highlighted that portable MRI led to changes in patient management in five cases. Importantly, portable MRI offers advantages in infection control as these devices can provide medical imaging to patients inside isolation wards, thus removing the need to transport infectious patients throughout the hospital.

Additional studies have characterized midline shift (MLS) in patients with ischemic stroke or hemorrhage on the 0.064T system (160,161). In a cohort of 102 patients, very-low-field MRI had a 93% sensitivity and 96% specificity for detecting the presence of MLS compared to SOC imaging based on manual identification of midline structures (160). In a follow-up study, an AI-based method (BrainInsight, Hyperfine, Guilford, CT) for assessing MLS was non-inferior to neuroradiologists (p<0.00001) (161). The MLS biomarker assessment is integrated in the scanner’s...
picture archiving and communication system (PACS), allowing for integration into the critical care workflow. Automated, quantitative biomarkers in lower-field, point of care (POC) imaging have the potential to reduce neuroradiological burden and may extend services to sites where radiologists are not readily available, provided findings are actionable in that context. **Figure 16B** illustrates an example of the midline shift function in a hydrocephalus patient imaged at our center.

In cases of stroke, time is brain (162). Mobile stroke units (MSU) were developed in 2015 to deliver appropriate therapy as quickly as possible (163). These vehicles are similar to traditional ambulances, but equipped with POC lab testing, a CT scanner, and personnel trained in stroke therapy. A recent study with over 1000 patients found that MSUs with onboard CT improved patient disability outcomes, led to a 34% reduction in time from stroke onset to t-PA administration (from 108 to 72 minutes), and decreased mortality rate from 11.9% to 8.9% (164). While MRI has not previously been integrated into MSUs because of siting and cost issues, new portable low-field MRI systems have been placed in vehicles for remote imaging (165,166). MSUs combined with portable MRI could offer rapid stroke imaging with the high tissue contrast and safety benefits of MRI.

While stroke imaging on portable systems seems promising, there are several outstanding challenges that need to be addressed. In a recent study, 2 neurosurgeons and a neuroradiologist rated image quality on the portable MRI system on a 5-point scale, 5 being the lowest quality: FLAIR (2.19 ± 0.98), T1-weighted (T1w) (2.6 ± 0.98), T2-weighted (T2w)-axial (2.47 ± 0.99), and T2w-coronal (2.88 ± 0.99) sequences had similar quality ratings with approximately 85% of images being deemed adequate for interpretation. However, only 27% of DWI were sufficient for interpretation and images had a correspondingly lower average quality rating (4.13 ± 1.02) (167). This highlights that DWI, which is the gold-standard for stroke imaging, remains challenging at
very-low-field. However, Sheth et al. noted improved image quality as newer software and hardware versions were released throughout their data collection (157).

Figure 17 - Images from a 27-year-old patient treated with ferumoxytol for iron deficiency anemia. (A) Ferumoxytol is a high-relaxivity contrast agent, which causes venous structures to be hyperintense on T1 and hypointense on T2 and T2-FLAIR. (B) The same patient was imaged with a magnetic resonance angiography (MRA) sequence (TE=5.17 ms, TR=10.34 ms, TI=0 ms, average = 1, resolution 1.5x1.5x3) to visualize arterial structures, including the internal carotid artery (ICA) and middle cerebral artery (MCA).

The breadth of sequences currently available on very-low-field scanners is another limitation. Only FLAIR, T1w, T2w, and DWI sequences have been described, and additional sequences need to be developed and optimized, such as perfusion and magnetic resonance angiography (MRA). Therefore, the evaluation of the infarct core, penumbra, and collateral flow remains deficient on these scanners. In our own work collaborating with a scanner manufacturer,
we used a high-relaxivity contrast agent (ferumoxytol) to perform contrast-enhanced MRA on a low-field system (Figure 17). Using this contrast-enhanced imaging technique, we were able to identify key venous (transverse sinus, sagittal sinus, and jugular vein) and arterial (carotid, middle cerebral, and anterior cerebral arteries) structures.

A final challenge for the adoption of lower-field imaging will be integration into critical care and neuroradiology workflows. Recent work has described protocols for integrating portable, very-low-field devices into ICU, emergency department (ED), and COVID-19 care settings (159,167). Neuroradiologists may require training on lower resolution images to become accustomed to the appearance of pathology at low field. Mazurek et al. illustrated several false negative cases where pathology could be seen upon closer examination, which indicates room for improvement as neuroradiologists become more familiar with lower-field images (157). For more details on low-field MRI and stroke imaging, the authors recommend the excellent review by Bhat et al. (146).

Promise

- Very-low-field MRI can assess stroke with relatively high sensitivity and specificity.
- Automated diagnostic tools are being integrated into portable MRI workflow.
- Portability could enable mobile stroke units equipped with MRI.

Challenges

- DWI sequences are the gold standard for stroke MRI, and presently DWI has lower quality on portable scanners.
- The ability to distinguish key features of stroke imaging, such as infarct core and collateral flow, remains limited.
- Perfusion-diffusion mismatch cannot currently be assessed in very-low-field scanners.
• Reading lower-field scans may require retraining for neuroradiology personnel.

4.4 Outpatient Neuroimaging

While most applications for portable MRI have centered on the neuro ICU, evidence is emerging regarding their efficacy for outpatient neurology use cases. Neurological disorders affect one billion people worldwide (3). They are the leading cause of disability and the second leading cause of death, killing an estimated 9 million people annually (4). In this section, we review emerging clinical evidence for lower-field devices in this setting and discuss how reduced costs and increased portability could impact when and where patients receive neuroimaging.

Lower-field MRI may play a role in management of multiple neurological conditions. Multiple sclerosis (MS) is a demyelinating disease affecting the brain and spinal cord, and MRI plays a vital role in guiding therapy decisions. Mateen et al. used an 80mT portable scanner to visualize demyelinating lesions in two patients with MS (168). They also noted cortical atrophy in one patient with advanced disease, a finding that has applicability in other neurodegenerative diseases, such as dementia. Some scanners have been specifically designed for imaging children with hydrocephalus, a condition in which the dominant fluid-filled spaces in the brain, the ventricles, become enlarged (169,170).

At our center, research has focused on adult outpatient neuroimaging applications, including hydrocephalus and multiple sclerosis. Our primary approach has been to collect same-day SOC and 64mT portable MRI imaging (Figure 18). We then compare modalities using manual ratings from clinicians and automated segmentation algorithms. In recent work, we probed device
sensitivity to MS lesions in 36 patients using paired 3T and 64mT data (171). We found the 64mT device was sensitive to white matter lesions and that radiologists as well as automated algorithms could detect lesions. We have also collected paired SOC and very-low-field data from 22 adult hydrocephalus patients and recorded programmable shunt settings before and after scanning (172). Radiologists were able to accurately categorize patient ventricle sizes as small, large, mixed, or within normal limits in all images and we determined that programmable shunt settings were altered by the very-low-field MRI.

In recent decades, researchers have identified numerous high-field MRI biomarkers for neurodegenerative and psychiatric disorders (173–176). Disease-specific patterns of brain atrophy

![Figure 18 - Patients imaged with 3T (top) and 64mT (bottom) at our center. A) T2w imaging of 28-year-old female with a history of traumatic brain injury and hydrocephalus, B) T1w imaging of a 31-year-old male with a history of seizure and bilateral hippocampal sclerosis, C) FLAIR imaging of a 22-year-old male with a pineal germ cell tumor, D) FLAIR imaging of a 66-year-old female with relapsing-remitting multiple sclerosis.](image-url)
can be evident on imaging prior to symptom onset, such as hippocampal volume loss in Alzheimer’s disease (177) or cortical atrophy in frontotemporal dementia (177,178). The presence of other features, such as white matter hyperintensities (WMH), can further aid in the differential diagnosis (179). In addition to providing diagnostic information, imaging biomarkers can serve as endpoints for clinical trials (180). The lower image resolution associated with lower-field devices has proven problematic for some conventional neuroimaging pipelines (181). To address this, Iglesias et al. recently developed a super-resolution algorithm (SynthSR) that takes lower resolution images and synthesizes a 1 mm isotropic T1w MPRAGE, which can be used for subsequent image processing. Using this approach, the group has demonstrated high correlation for key brain regions (i.e. hippocampus, thalamus, ventricles, cortical gray matter, etc.) between 3T and SynthSR-enhanced very-low-field images (127,182). Deoni et al. have also demonstrated the ability to generate 1.5 mm isotropic T2w images by registering and averaging three orthogonal slice plane acquisitions (126). Although it remains to be seen if high-field biomarkers can be validated on lower-field imaging, these initial super-resolution results provide a promising avenue.

MRI resources tend to be concentrated in population centers, which has resulted in reduced imaging access in rural areas and introduced sampling bias into many clinical neuroimaging research studies (183). This has obvious consequences for the diagnosis and treatment of individuals in areas with limited MRI resources. To increase imaging access, tractor trailers have been retrofit with 1.5T magnets. These mobile scanners can enable device cost sharing between multiple hospitals and have permitted imaging in restricted populations (184). However, these devices cost millions of dollars and have complicated infrastructure, which has limited their deployment. Recently, research groups have been experimenting with retrofitting vans with permanent magnet lower-field devices. Nakagomi et al. placed a 0.2T low-field magnet in a Mercedes minivan for mobile extremity imaging (166). They envisioned the device being deployed
to sporting events or areas without MRI access. Deoni et al. similarly retrofitted a Ford transit van with a 0.064T very-low-field magnet and demonstrated neuroimaging in both pediatric and adult patients at the participants’ homes (165). The estimated cost of the project was 110,000 USD, a fraction of cost to purchase a mobile 1.5T tractor trailer.

Increased healthcare costs in the United States have led to a critical evaluation of our spending, with medical imaging being identified as a contributing factor (185). In addition to optimizing current imaging practices (186), increased reliance on lower-field devices may offer a cost-effective means of enhancing MRI value as we transition from volume-based to value-based care systems. The highest concentration of MRI devices worldwide is in Japan where MRI reimbursement rates are identical across field strengths (9,187). This has led to widespread adoption of mid-to-low-field devices, which offer increased profitability because of their lower cost per examination. While there are undoubtedly cases where high-field MRI is more clinically appropriate, it may be reasonable to adopt a similar approach to Japan, where high-field scanners are concentrated in healthcare centers and mid-to-low-field devices are more widely available. While SNR is proportional to field strength, this may not be the best metric for determining how much value different MRI systems contribute to diagnostic accuracy, patient outcomes, and societal benefit (188). Using lower-field devices, it may be possible to triage patients and reduce scheduling demands on high-field scanners, which could result in decreased diagnostic delays and increased patient satisfaction (189,190).

Although lower-field devices may augment SOC imaging in high income countries, they are likely to play a more impactful role in LMICs. In 2016, an estimated 84 MRI units serviced West Africa, an area with a population of over 370 million (191). For comparison, in 2019 the United States had over 13,000 devices servicing a smaller population (328 million). Mid-to-low-field devices already play a dominant role in MRI services in West African countries, with the vast
majority (77.6%) of devices in Nigeria being low-field strength (<0.3T) (191,192). Neuroimaging is the dominant application, with one center reporting that over 90% of studies requested were for brain (49.9%) and spine (45.6%) imaging (193). However, given that the average MRI cost is ~500 USD, that many patients live of less than a 1 USD per day, and that services are typically paid by patients out-of-pocket, even mid-to-low-field scanners are beyond the reach of a significant portion of the population (193).

A primary motivator for the development of ultra-low-field and very-low-field systems has been increased geographical and financial access (194). This includes devices targeting pediatric hydrocephalus, which has a high prevalence in Africa (169,170,195). These more targeted systems may be cheaper to produce and easier to service, allowing for lower out-of-pocket costs to patients. Ogbole et al. also noted their scanner experienced significant downtime due to lack of technical support or service materials (193). When designing devices for LMICs, special consideration must be given to available resources and technical expertise (196,197). Additionally, many LMICs have a dearth of radiologists and radiographers (198). In these cases, remote readers or automated algorithms may be able to provide diagnostic support, allowing countries to stretch scarce resources (199,200). Equally important to proper device design is the appropriate and equitable introduction of these devices into society. Recently, working groups of researchers, clinicians, MR vendors, and local stakeholders have convened to provide guidelines and address key ethical, legal, and social issues surrounding portable MRI (201,202). Continued engagement of these working groups will be essential for providing updated recommendations as new hardware and software are released. For more details on imaging accessibility, the authors recommend the excellent review by Geethanath et al. (194).

Promise
• Mobile MRI systems offer patients and researchers more imaging opportunities, expanding the definition of “outpatient.”
• Lower out-of-pocket costs could increase medical imaging access in LMICs and rural areas.
• Lower cost scanners could reduce medical imaging expenditure.
• Simpler lower-field systems, with less moving parts and lower cooling and power requirements, may be more robust and require less service and maintenance.

Challenges

• Portable lower-field systems may require changes to reimbursement structure.
• Many LMICs lack of trained radiologists and associated medical imaging specialists.
• Maintaining high ethical, legal, and social standards in a rapidly changing device landscape.

4.5 Intraoperative MRI and MRI-guided procedures

MRI has become an integral part of neurosurgery, allowing surgeons to plan procedures based on an individual’s anatomy and to monitor patients for complications. Perhaps the most important problem at the interface of neuroimaging and surgery is the accurate localization of structures. In the 1980’s, frame-based stereotaxy became the first widely used systematic method for localizing intracranial structures (203). These systems, which fix the patient’s head in a physical frame to relate a coordinate system, have largely been replaced by frameless neuronavigational systems, which utilize fiducial markers instead of a frame (204). Today, frameless neuronavigational systems are the most widely-deployed localization method used in high-income countries (205).
Despite their physical differences, both frame-based and frameless methods for neuronavigation rely on preoperative imaging, and significant anatomical distortions are known to occur as tissue is removed and cerebrospinal fluid is lost during surgery. These methods also do not permit intraoperative monitoring for complications, such as hemorrhage. To account for these shortcomings, researchers in the late 1990s began developing intraoperative MRI (iMRI) approaches using mid-to-low-field scanners. Two initial iMRI approaches were developed: the Boston concept (206), in which the surgery is performed in the scanner, and the Heidelberg concept (207), in which the patient is transported to a nearby scanner. Later, the idea of bringing the scanner to the patient was explored using both high-field (208) and lower-field (209) scanners. Approaches where either the patient or the scanner are moved have become the most widely adopted because they permit higher field-strength magnets, unrestricted access to the patient, and the use of traditional ferromagnetic surgical implements (210).

iMRI has largely been pioneered in neurosurgery and is most widely used for brain tumor resections, where the superior soft tissue contrast and 3D visualization facilitates the goals of maximal tumor resection, minimal healthy tissue removal, and monitoring for surgical complications such as hemorrhage. A wide range of studies have demonstrated that iMRI increases the gross total resection rate and improves patient outcomes (211). Other common uses for iMRI include accounting for brain shift during surgery (212), needle guidance for biopsy (213), functional MR guidance to avoid eloquent cortex (214), tractography to avoid major white matter tracts (215), guidance and temperature monitoring during thermal ablations (216), seizure focus resection (217), and placement of intracranial devices such as deep brain stimulators or stereoelectroencephalography probes (218).

Several innovative approaches to mid-to-low-field iMRI have been developed, including the original 0.5T Signa SP (GE) pioneered by Black et al. (206), the 0.2T Magnetom Open
(Siemens) pioneered by Tronnier et al. where patients were moved intraoperatively (207), and the semi-portable 0.12-0.3T PoleStar N-10, N-20, and N-30 systems (Odin Medical, later Medtronic) (209). Glioma resections constitute a large portion of studies using these devices (211,219,220), although other applications include pituitary tumor resection, epilepsy resection, and skull-base tumor resections. While studies using these devices demonstrated improvements over standard surgery, including increases in gross total resection (221), remission rate (222,223), and survival time (219,220), tumor remnants found using high-field imaging were also reported (223–225).

While a range of field strengths have been explored for iMRI, today most devices are either 1.5T or 3T. Higher-field strengths have been favored because they permit higher image quality with faster acquisition times (226), allow for a greater resection extent (227,228), and have increased sensitivity to enhancing neoplasm (226,229), which is the main predictor of surgical outcome. High-field systems have demonstrated clinical and economic benefits (230), while evidence for low-field systems has been mixed (231). High-field devices have disadvantages associated with their increased magnetic field strength, including increased susceptibility artifact, hardware interactions, shielding requirements, a larger 5-gauss line, increased safety precautions, and MR compatibility issues (226). Additionally, iMRI systems traditionally require a significant capital investment (232), more staff training (232), and longer procedure times (210), all of which have likely slowed adoption.

Newer lower-field devices may overcome some of the disadvantages of high-field scanners while improving upon the shortcomings of prior lower-field iterations. With a sufficiently low magnetic field, surgeries can be performed using traditional implements and without moving the patient or scanner. Additionally, staff training burden and safety precautions are reduced. Some researchers have advocated for the development of mid-field systems equipped with the latest technology developed for high-field systems (121,122,154,155). Campbell-Washburn et al.
modified a 1.5T Siemens Magnetom Aera to operate at 0.55T while maintaining gradient performance and using a 16-channel head coil (122). A similar design is now commercially available as the 0.55T Siemens Magnetom Free.Max. Campbell-Washburn et al. described a cohort of seven patients that underwent successful MRI-guided right heart catheterization using the mid-field scanner and demonstrated reduced radiofrequency-induced heating in guidewires, catheters, and pacemakers previously deemed unsafe at 1.5T. Synaptive Medical offers the 0.5T Evry system, which was designed for iMRI applications. Initial reports have demonstrated high-quality imaging, including diffusion tensor imaging sequences, and reduced risk of radiofrequency-induced heating (233–235). To our knowledge, no publications have reported either scanner being used for iMRI neurosurgical applications to date, but the increased MR compatibility and potential for reduced scanner costs merit further investigation.

At lower field strengths, the 0.066T Promaxo system has been proposed as an point-of-care method for guiding prostate biopsies by urologists (236,237). Prior to the procedure, high-field imaging is collected to identify targets for biopsy. Very-low-field imaging is then collected during the procedure and registered to high-field imaging to provide needle guidance to the regions of interest. In phantom studies, the error in needle guidance was less than 3 mm on average. In early in-vivo reports, the MRI guided procedure increased sensitivity to prostate cancer by 37% over blinded systematic biopsy (MRI: 12/16, blinded systematic biopsy: 6/16). To date, there have yet to be reports of any modern lower-field MRI systems for neurosurgical applications. However, low-field devices from prior generations, such as the PoleStar system which is no longer available, had cost approximately 750K USD annually (231). Newer lower-cost scanners have the potential to change the cost-benefit analysis and may drive adoption of iMRI in areas where it was previously deemed cost-prohibitive (231).
Although new lower-field devices offer reduced scanner costs and have improved in image quality, there are still significant challenges. Most critically, there is the question of device sensitivity. Prior studies of low-field devices frequently reported remnant tumors on high-field follow-up. Another concern is that conventional gadolinium contrast is suboptimal at lower-field strength, and either dosage must be increased or higher relaxivity agents must be developed for its use to be feasible (238). Finally, lower-field systems often have a smaller field-of-view, which may not be suitable for certain surgeries, such as those involving the skull-base or pituitary fossa, where a transsphenoidal approach is commonly used (223–225). Configuration of the magnet and coil must be designed to accommodate surgical implements, monitoring equipment, and staff needs during surgery. For more details on iMRI applications, the authors recommend the review and book by Hall et al. (218,239).

Promise

- Low-cost systems could enable intraoperative MRI guidance in situations that were not previously economically feasible.
- Lower-field MRI could be used to check for surgical complications, such as hemorrhage, prior to ending the surgical procedure.
- Lower-field scanners have reduced requirements for radiofrequency shielding, operational safety, staff training, 5-gauss line distance, and MR compatibility for traditional surgical implements.

Challenges

- iMRI increases total procedure time and device costs are high, which may have discouraged adoption of previous lower-field iMRI devices.
• Previous lower-field devices have been supplanted by high-field devices because of superior image quality or scanner versatility.

• Conventional gadolinium contrast is imaged suboptimally at lower-field strengths. This could prove challenging for cases where resection of enhancing tissue is the primary predictor of patient outcome, such as glioblastoma.

• Laser ablations require precise spatial localization and real-time MR thermography, and temperature resolution is known to decrease with magnetic field strength.

• Lower-field systems have smaller fields-of-view and device configurations will need to be optimized to facilitate surgical access to the patient.

4.6 Pediatrics

Pediatric neuroimaging is increasingly used for both clinical and research purposes (240,241). There are unique challenges associated with imaging pediatric patients, including safety concerns, obstacles to image acquisition, and differences in how images are analyzed (242). These barriers can prevent the direct translation of research or clinical practices from adult to pediatric populations and highlight the importance of evaluating each step of the neuroimaging pipeline with the target population in mind. Here, we will highlight some of the features of pediatric neuroimaging and discuss how lower-field MRI could contribute to the process.

The dominant modalities for pediatric neuroimaging are ultrasound (US), CT, and MRI. US is primarily used to diagnose disorders such as hydrocephalus and intracranial hemorrhage in the first 6 months of life, prior to the closure of the anterior fontanelle (243,244). CT is commonly used for trauma-related neuroimaging, as it offers fast imaging with good contrast between blood, bone, and brain (244). Acquisition times are short, so sedation of pediatric patients is rarely
required. However, CT has less soft tissue contrast and exposes patients to ionizing radiation, making it less desirable for repeated imaging, such as for monitoring patients with shunted hydrocephalus (245–247). While MRI offers superior soft tissue contrast and does not use ionizing radiation, it is more expensive and less widely available than CT. This is particularly the case in rural areas and in many LMICs (248). MRI sequences are also significantly longer than CT, often necessitating patient sedation and posing additional safety risks (242). Finally, MRI hardware, such as the head coil, needs to be specialized for pediatric patients to provide optimal imaging (249).

Advantages of lower-field MRI that are appealing for pediatric applications include open scanner designs and reduced scanner noise. Patient compliance can be particularly challenging for children under 6 years old, who often require sedation (242). With open scanner designs in lower-field MRI, claustrophobia is less of a problem. The need for sedation may also be lower, as parents or providers can access children during the scan. Rupprecht et al. compared the sedation or anesthesia requirements of 274 pediatric patients imaged on a 0.2 tesla open bore MRI (Siemens Magnetom Open, Siemens) to 111 patients imaged on a standard high-field, closed bore system. In the open system, 27% (74/274) of patients required sedation compared to 47% (52/111) on the closed system, with the effect most pronounced in children under 10 years old. The authors also reported requiring lower sedation doses and that monitoring patients was easier. Moreover, lower-field devices have reduced acoustic noise, which is particularly useful for infants, who are often asleep during scans and may need ear protection (242,251).

Smaller or portable scanners can also be integrated into neonatal or pediatric ICUs and in operating rooms. Low-field systems have been deployed for thoracic, orthopedic, and neurosurgical applications (252–254). Low-field systems have also been integrated into pediatric and neonatal ICU settings, permitting neuroimaging for patients who cannot be transported to fixed MRI scanners (255,256). In the early 2000s, Whitby et al. compared the SOC imaging (US) to a 0.17T
(InnerVision MRI Ltd, London, UK) in 134 (89 controls & 43 with suspected pathology) neonatal ICU patients (255). In 56% (24/43) of patients, the MRI provided additional clinical information above what was found using US. In 40% (17/43) of cases, the US read was normal, while the MRI detected five cases of subdural hematomas, five cases of hypoxic ischemia, and seven additional findings. Whitby et al. noted the device cost (~£150K in the United Kingdom) and relative cost per treatment (~£60 in the UK) were similar between US and MRI (255). More recently, Aspect Imaging received FDA clearance for their 1T mid-field Embrace neonatal MRI system, which is designed to be embedded in the NICU. Thiim et al. compared the 1T system to US in the NICU and conventional 3T scanning with patient transport outside the NICU. Compared to US, the 1T scanner provided significant clinical benefit, with abnormal findings identified in 15 additional cases (1T MRI: 59, US: 44). The authors also reported greater sensitivity to white matter injury (1T MRI: 17, US:7), hypoxic ischemia (1T MRI: 2, US:0), and hemorrhage (1T MRI: 25, US: 20) (256). For 32 patients, 3T comparison imaging was collected. Reports of brain injury were largely concordant between 3T and 1T, with two notable exceptions: one case of punctate foci of susceptibility noted at 3T but not 1T and one case of polymicrogyria noted at 1T but not 3T.

Portable very-low-field systems have been applied to pediatric neuroimaging to track neurodevelopment. Deoni et al. described the first use of a portable MRI on a cohort of 42 healthy children (age range of 6 weeks to 16 years) imaged using a 0.064T scanner (181). The researchers calculated brain volume estimates for gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) and replicated known developmental trajectories from 3T imaging studies. The authors reported greater scan success rate with 0.064T (89%) than 3T (75%), presumably related to the open bore design. However, researchers noted that some sequences and analytical software may require optimization to work in pediatric populations. The team’s volumetric analyses were performed on T2w sequences, which offered superior anatomical contrast
on the very-low-field scanner, instead of the T1w sequences typically used in morphometrics. White matter myelination changes rapidly during childhood, and developing brains have higher water content in unmyelinated white matter. Both water content and magnetic field strength influence the relaxivity properties of the tissue (249,257). The researchers used sequences designed for imaging adult patients, which likely need to be optimized based on patient age to provide better T1 tissue contrast in pediatric patients. Additionally, researchers utilized two common neuroimaging analysis software packages, ANTs (45) and FreeSurfer (258). They noted that FreeSurfer failed to process the low-resolution data accurately, although this may also improve with sequence optimization. Finally, the researchers only replicated group-level developmental trajectories, and it remains to be seen if accurate volume estimates can be obtained at the individual level.

The Deoni et al. study highlights how expanded access to portable, lower-cost scanners could change how large neuroimaging studies are conducted, shifting focus away from academic hospitals in HICs to including more children from rural areas and LMICs. Lower-cost scanners could also enable larger sample sizes in research by reducing the cost per patient scanned. Recently, a group at the Queen Elizabeth Central Hospital in Malawi recently reported scanning 260 patients with a 64mT portable MRI, including examples of subdural empyema, a collection of purulent material around the brain, and malarial encephalopathy (259). Research applying very-low-field MRI in LMIC is likely to increase in coming years with increasing investment in the field.

Despite the many advantages lower-field MRI offers for pediatric neuroimaging, there are significant barriers that must be overcome before the technology can be widely adopted. Foremost, sequences, hardware, and analytic software need to be optimized for pediatric populations. Multiple researchers have reported that T1w and DWI sequences require quality improvements, some specifically referencing optimization for pediatric patients (159,167,181,260). Many current lower-
field systems contain hardware designed for adults, which may require alterations for pediatric patients (249). Additionally, some sequences used for neurodevelopmental research are difficult to obtain on lower-field systems, including functional, perfusion, and high-angle-diffusion sequences. Finally, because of the lower SNR associated with a lower magnetic field, lower-field sequences are often longer, which reduces patient compliance and exacerbates motion issues.

Promise

- Open MRI designs offer easier access during pediatric neuroimaging, allowing parents or professionals to care for and comfort the child or permit usage of medical treatment devices during imaging.
- Infants are often imaged while asleep to minimize motion, and weaker gradients offers a significant reduction in scanner noise.
- Lower-field scanners provide a high contrast, non-ionizing radiation method for assessing chronic neurologic conditions, such as hydrocephalus.
- Lower-field scanners may have clinical (intensive care, surgical, as well as outpatient diagnostics and monitoring) and research (neurodevelopmental) applications in pediatric populations.
- Lower scanner costs could increase access to imaging in rural areas and LMICs as well as permit acquisition of large neuroimaging studies.

Challenges

- Lower-field sequences and hardware (such as head coils) need to be optimized for pediatric patients, with T1w and DWI sequences requiring the most improvement.
- Lack of some sequences used in neurodevelopmental research, most notably functional, perfusion, and high-angle diffusion sequences.
 Pediatric sequences and scan sessions are often shorter to combat patient motion; lower-field sequences typically have longer acquisition times.

 It remains unknown if brain volumetrics can be reliably calculated at the individual level for pediatric patients.

### 4.7 Orthopedics and musculoskeletal imaging

Orthopedics and musculoskeletal imaging (MSK) were early adopters of low-field MRI and remain one of the few specialties where low-field is used in clinical practice today. Lower-field strength scanners are particularly attractive to orthopedics because metal implants are common in their patient population and unique scanner designs (such as open bores, extremity specific scanners, and weight-bearing scanners) offer more clinically tailored imaging (261). Several manufacturers, including FONAR, Esaote, and Paramed, currently offer low-field MRI scanners targeted at orthopedic and spinal applications (262,263). These devices focus on minimizing imaging costs or feature rotating tables which permit weight-bearing and kinematic imaging. Unlike conventional closed-bore systems, extremity specific and open bore scanner configurations permit central positioning of limbs in the magnetic field, which increases image quality (261). Although these advantages have given low-field scanners some traction in orthopedics, they remain relatively uncommon in clinical practice today (263).

Although low-field MRI offers a cost-effective means for orthopedic imaging (264), there remain significant challenges that must be overcome before these scanners can become more widely adopted. The three primary obstacles are user perceptions of lower-field image quality, developing the full range of clinically necessary sequences, and loss of quality control by radiologists. User perception of image quality and its impact on diagnostic value is perhaps the
most important problem facing lower-field MRI. In the 1990’s and early 2000’s, low-field MRI scanners were directly compared to standard high-field systems across a range of orthopedic applications. In shoulder imaging, most studies found equivalent performance between scanners (265–268) with the notable exception of Magee et al., who reported that subsequent high-field scans changed reviewer interpretations in 9/40 patients (269). However, as noted by Thomsen et al., their results were for a single low-field system and may not generalize to low-field devices more broadly (270). In knee imaging, most studies reported equivalence (271,272), although a meta-analysis including 29 knee MRI studies identified a significant reduction in diagnostic performance for ACL tears (273). Low-field scanners were evaluated for numerous other orthopedic applications, including the elbow (274), hand (275), and foot (276). While many studies reported equivalent diagnostic performance between low-field and high-field scanners, there remained significant concern that the lower SNR would translate to missed diagnoses and perhaps legal liability (272).

In the first report describing knee scans on a commercial 64mT very-low-field scanner, raters visualized the quadriceps tendon, patellar tendon, and posterior cruciate ligament (PCL) easily, but had more difficulty with the anterior cruciate ligament (ACL), iliotibial band, medial collateral ligament (MCL), and lateral collateral ligament (LCL) (277). Some sequences may need optimization to provide sufficient visualization of key knee anatomy. Reduced image artifacts surrounding implants could be a substantial benefit when using lower-field devices. In the mid-field range, Khodarahmi et al. found a 45-64% reduction in artifact when using 0.55T compared to 1.5T to scan patients with implants, even when 1.5T scanners used a slice encoding for metal artifact correction (SEMAC) protocol (278). By comparison, the 17-28% reduction in SNR was relatively modest and still provided sufficient image quality for interpretation.

Another challenge at lower-field strengths is replicating the sequences that radiologists have come to rely on at high-field. In orthopedics, fat suppression techniques are frequently used
to increase contrast when evaluating cartilage, menisci, or bone marrow (279). However, many fat suppression techniques rely on the chemical shift between lipids and water, which is proportional to magnetic field strength (280). Recently Bellisari et al. demonstrated a fat suppression technique on a 0.25T low-field scanner and found comparable diagnostic performance to 1.5T (281). Although this shows promise, this technique may not transfer to even lower field strengths and represents only a single sequence of the multiple sequences that still need to be optimized.

Finally, as scanners become increasingly used by non-radiologists and in non-traditional health care settings, there is significant concern that radiologists will not be able to oversee quality control of imaging or be available for image interpretation. Portable, very-low-field devices could be deployed in non-traditional settings, such as in a field hospital or the sidelines of a sports arena, for musculoskeletal imaging. Several groups have built open source Halbach array scanners, including Guallart-Naval et al., who reported the use of one such 70mT Halbach device to scan a patient with a knee implant in various scanning environments (135). They compared SNR when scanning in a patient indoors versus outdoors and when powering the device using a wall outlet versus a gas generator. SNR was most affected in the outdoor setting when powered by a gas generator, although even in this setting SNR was considered acceptable by the authors. Additionally, there was minimal artifact near the implant in all imaging settings.

The idea of providing low-field scanners in remote settings with diagnostic support tools has been gaining traction. Raman et al. used a neural network to classify knee effusion in simulated lower-field images (282). Their network had greater accuracy than radiologists (47.2% vs 41.7%), indicating that such technology could provide diagnostic feedback in the event a physician is not readily available. Automated methods for segmenting knee anatomy at low-field also show promise for providing diagnostic support (283). While this could further be mitigated by telehealth, it remains unclear how health systems will adapt to increased imaging outside the radiologist realm.
For more details on lower-field MRI and orthopedic applications, the authors recommend Ghazinoor et al. (272).

Promise

- Cheaper scanners can be tailored to specific applications (hand/foot/limbs) and could be integrated into orthopedic departments.
- Unique scanner designs allow expanded patient positioning and mobility and enabling weight-bearing and kinematic studies.
- Open designs permit positioning of limbs in isocenter.
- Portable scanners could offer unique opportunities for sports medicine.
- Metal artifact is reduced as lower-field strengths (orthopedics has above average device/implant rates).

Challenges

- Reduced sensitivity to certain anatomic structures.
- Less spectral separation between water and fat, rendering common fat-suppression techniques inadequate (though promising DIXON results).
- Loss of quality control by radiologists.
Chapter 5 - Simulated diagnostic performance of low-field MRI

As novel imaging devices are introduced there is an inherent paucity of data available to evaluate their efficacy across the full range of pathology encountered in the clinical setting. In this chapter, we have proposed a method for guiding evaluation of novel imaging devices via simulated trials. We apply domain transformations to publicly available datasets to generate simulations with similar image quality to the device under investigation. We then leverage machine learning to perform automated pathology detection as a surrogate for traditional reads by a radiologist. In this manner, we can investigate the potential sensitivity of a new device to a range of pathologies, without costly and time-consuming prospective data collection. This work was published as a preprint on medRxiv (284) and as an article in Magnetic Resonance Imaging (285), with T. Campbell Arnold as co-first author.

Simulated Diagnostic Performance of Low-Field MRI: Harnessing Open-Access Datasets to Evaluate Novel Devices

T. Campbell Arnold¹,² *, Steven N. Baldassano¹,² *, Brian Litt¹-³, Joel M. Stein¹,²,⁴

1. Department of Bioengineering, University of Pennsylvania
2. Center for Neuroengineering and Therapeutics, University of Pennsylvania
3. Department of Neurology, University of Pennsylvania
4. Department of Radiology, University of Pennsylvania

* These authors contributed equally to this work.
Abstract

The purpose of this study is to demonstrate a method for virtually evaluating novel imaging devices using machine learning and open-access datasets, here applied to a new, low-field strength portable 64mT MRI device. Paired 3T and 64mT brain images were used to develop and validate a transformation converting standard clinical images to low-field quality images. Separately, 3T images were aggregated from open-source databases spanning four neuropathologies: low-grade glioma (LGG, N=76), high-grade glioma (HGG, N=259), stroke (N=28), and multiple sclerosis (MS, N=20). The transformation method was then applied to the open-source data to generate simulated low-field images for each pathology. Convolutional neural networks (DenseNet-121) were trained to detect pathology in axial slices from either 3T or simulated 64 mT images, and their relative performance was compared to characterize the potential diagnostic capabilities of low-field imaging. Algorithm performance was measured using area under the receiver operating characteristic curve. Across all cohorts, pathology detection was similar between 3T and simulated 64mT images (LGG: 0.97 vs. 0.98; HGG: 0.96 vs. 0.95; stroke: 0.94 vs. 0.94; MS: 0.90 vs 0.87). Pathology detection was further characterized as a function of lesion size, intensity, and contrast. Simulated images showed decreasing sensitivity for lesions smaller than 4 cm$^2$. While simulations cannot replace prospective trials during the evaluation of medical devices, they can provide guidance and justification for prospective studies. Simulated data derived from open-source imaging databases may facilitate testing and validation of new imaging devices.

5.1 Introduction

Modern medical imaging has become a mainstay of optimal patient care, particularly in the diagnosis and management of patients with neurologic disease. While the availability of imaging technology has dramatically increased worldwide in recent decades, the expense and operational
complexity of standard imaging systems limits access in underserved areas and developing countries (191). This so-called “radiology divide” leaves about ninety percent of the world’s population without access to magnetic resonance imaging (MRI) (286) and almost two-thirds of the population without even basic imaging technology such as ultrasound and X-ray radiography (10–12).

Low-field (LF) strength MRI systems aim to make MRI more accessible, promising lower cost, portability, fewer magnetic field-related safety concerns, and ease of use (122). Such devices could decrease healthcare expenditures, improve availability in underserved areas, and provide a convenient and ionizing-radiation-free modality for routine or monitoring studies. Portable LF MRI systems may be suitable for hospitalized patients, such as those in intensive care units or isolation wards, for whom transport to a standard clinical scanner carries unacceptable risk (157,287–289). More broadly, portable LF MRI units could potentially be used in ambulances, emergency departments, physician’s offices or rural clinics (181,201).

While LF MRI presents clear practical advantages, these systems produce images with lower signal-to-noise, resolution, and tissue contrast compared to their high-field strength (HF) counterparts and are largely designed to complement, and not replace, standard MRI. Prior to deployment for clinical use, the diagnostic capabilities of novel imaging technologies such as portable LF MRI should be evaluated across a wide range of patients and pathologies. The standard approach for device evaluation, improvement, and optimization requires recruiting large numbers of patients and manual image review by radiologists. This process is costly and time-consuming, which can limit the device development cycle. Moreover, selecting target use cases is difficult without basic information about device sensitivity. A complementary approach is to simulate LF images from existing, publicly available HF datasets and leverage machine learning for image interpretation to guide prospective clinical study design. Such datasets, typically compiled for
machine learning competitions (43,290–292) or collaborative research programs (293,294), span broad ranges of pathology and offer a wealth of information for retrospective analysis (295).

Here, we propose a generalizable method for image simulation and interpretation that can guide prospective clinical studies of novel neuroimaging devices. We employ a simple empiric method to transform existing images to a custom domain, converting high-resolution 3T MR images aggregated from several open-access databases to images matching the resolution and quality of those acquired on a portable 64mT LF MRI scanner. Quantitative measures and manual ratings by radiologists are used to compare image quality between real and simulated 64mT imaging. Separately, convolutional neural networks are trained to detect pathology in axial slices from the HF or simulated LF images, and detection performance is compared between image pairs to characterize the potential diagnostic capabilities of 64mT LF MRI. While automated lesion detection in simulated images does not guarantee detection on actual devices, simulated performance may help indicate whether pursuing a prospective study for a given application is promising. Applied here to LF MRI, this simulated trial approach offers a broadly applicable means for evaluating and optimizing novel medical imaging technologies to complement traditional imaging clinical trials.

5.2 Materials & Methods

5.2.1 Data collection

Paired portable 64mT LF (Hyperfine) and same-day standard clinical 3T HF (Siemens) brain MRI data were collected as part of an ongoing research study approved by the University of Pennsylvania Institutional Review Board. Participants provided informed consent. To develop the domain transformation, we used data from six adult patients with known or suspected
hydrocephalus. To validate and assess the generalizability of the domain transformation, we used data from ten adult patients with multiple sclerosis (MS). Fluid-attenuated inversion recovery (FLAIR) images covering the whole brain were collected on each scanner. This sequence was chosen because it 1) is fundamental to clinical imaging for each type of pathology detailed below, 2) provides the most lesion conspicuity across pathologies, 3) is more robust at low field than other sequences such as diffusion-weighted imaging (DWI), and 4) does not rely on exogenous contrast mechanisms. For 64mT imaging, patients received the following 3D fast spin-echo scan optimized for typical brain tissue contrasts (TE=200 ms, TR=4 s, TI=1.4 s, averages=1, scan time=9:29 min, resolution=1.6x1.6x5 mm). For 3T imaging, hydrocephalus patients received clinical axial 2D FLAIR imaging with variable sequence parameters (TE=96-141 ms, TR=8-10 s, TI=2.2-2.55 s, resolution=0.72-0.94 x 0.72-0.94 x 3-6 mm), while MS patients received a standardized 3D FLAIR sequence (TE=398 ms, TR=5 s, TI=1.6 s, averages=1, scan time=5:02 min, resolution=1mm isotropic).

Separately, retrospective axial FLAIR images obtained at 3T for a range of pathologies were aggregated from several open-access sources (43,290–292,296). Pathologies consisted of high-grade glioma (HGG, N=259), low-grade glioma (LGG, N=76), stroke (N=28), and MS (N=20) (43,290–292). Each dataset contained manual segmentations of lesions, which generally manifest as hyperintense areas on FLAIR imaging. These datasets incorporate a range of lesion sizes and signal intensities that can be quantified using the provided lesion segmentations. HGG and LGG lesion segmentations on FLAIR imaging include areas that may represent vasogenic edema or non-enhancing infiltrative neoplasm as well as enhancing components when present. Additionally, non-lesional control scans (N=5) were drawn from the OASIS3 dataset (296). Table 4 contains information about the different public datasets used in this study. Related web addresses as of publication are listed in section 5.2.9 Code and data availability.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pathology</td>
<td>Glioma (HGG &amp; LGG)</td>
<td>Stroke</td>
<td>Multiple Sclerosis (MS)</td>
<td>Multiple Sclerosis (MS)</td>
</tr>
<tr>
<td>Patients</td>
<td>335 (259 HGG, 76 LGG)</td>
<td>28</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Center / Scanners</td>
<td>19</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Classes</td>
<td>3 (enhanced, non-enhanced, edema)</td>
<td>1 (infarct)</td>
<td>1 (MS lesion)</td>
<td>1 (MS lesion)</td>
</tr>
</tbody>
</table>

Table 4 – List of open-access neuropathology datasets used in this study. Abbreviations: High Grade Glioma (HGG), Low Grade Glioma (LGG), Multiple Sclerosis (MS).

5.2.2 Domain transformation: High-field to low-field MRI

To generate simulated LF images from HF data, we employed a simple image transformation using 3T and 64mT image pairs from three of the hydrocephalus patients. Transformation steps are listed in Figure 19A and include registration, brain extraction, re-slicing, Gaussian smoothing, and noise filtering. After registration, brain extraction, and re-slicing simulated images match LF resolution; however, SNR remains substantially higher than actual LF imaging. To better match LF image quality, a series of smoothing kernels and noise filters were applied to simulated images. Noise filter amplitude and smoothing kernel standard deviation were parameterized and fit using training data. First, random noise was smoothed using a 3-D Gaussian kernel with a 0.5 standard deviation. Amplitude of the noise filter was parameterized and added to the image. Next, the image was smoothed using a 3-D Gaussian kernel with a parameterized standard deviation. Finally, an additional noise filter was applied with a parameterized amplitude and smoothing kernel. To determine optimal parameters, we minimized the difference in histogram features between real and simulated images. The objective function consisted of the first three statistical moments (mean, standard deviation, and skewness). An example HF/LF pair and the simulated LF image can be seen in Figure 19B with the matched intensity histograms shown in
Figure 19C. The transformation was applied to the HF images collected from open-access datasets to produce simulated LF images for gliomas, stroke, and multiple sclerosis.

Figure 19 - Generating simulated low-field (LF) images. (A) Steps in the image processing pipeline from the original 3T image (upper-left) to the simulated LF version (center). (B) Example skull-stripped axial FLAR images from a clinical 3T (left) and a 64mT LF MRI scanner (center). The 3T image was passed through the image transformation pipeline to produce the simulated LF image (right). (C) To generate simulated LF images resembling actual LF images, histogram features (mean, standard deviation, and skewness) were used to guide image transformation. The intensity histogram distributions relative to actual 64 mT images are shown before (left) and after (right) transformation.

5.2.3 Domain transformation: Quantitative validation

A quantitative validation of the domain transformation was performed using data from three additional hydrocephalus patients. The image transformation method was quantitatively assessed using the gradient entropy, $F$, as a measure of perceived diagnostic image quality. This metric was derived by McGee et al. (297):

$$F = - \sum_{i,j} h_{i,j} \log_2 [h_{i,j}],$$  \hspace{1cm} (1)
where $g_{i,j}$ is the pixel value at coordinate $i,j$ and $*$ represents the convolution operation. Of 24 metrics evaluated, McGee et al. found gradient entropy (Equation 1) to have the strongest correlation with radiologists’ perception of image quality in structural MRI. We compared gradient entropy values between 3T, real 64mT, and simulated 64mT images using a paired t-test.

\[
    h_{i,j} = \frac{||[1-1]^*g_{i,j}||}{\sum_{ij} ||[1-1]^*g_{i,j}||},
\]

### 5.2.4 Domain transformation: Radiologist validation

For the domain transformation approach to be valid, real and simulated LF images should have similar image quality and interpretability for radiologists. To assess the diagnostic quality of images, we compared ratings from three neuroradiologists (with 4, 5 and 9 years in clinical practice) for real 64mT, simulated 64mT, and 3T imaging from ten MS patients (Figure 20B). All imaging was coregistered and paired axial slices were drawn from each image type. We asked the neuroradiologists to rate slices from each image type using a 5-point Likert scale (298). Each neuroradiologist rated a total of 90 image slices and slice order was randomized. For each slice, neuroradiologists were asked: 1) Do you see white matter lesions in this image? (yes/no), 2) How confident are you in this rating? (1 = excellent, 2 = good, 3 = average, 4 = poor, and 5 = random), 3) What is the diagnostic quality of the image? (1 = excellent, 2 = good, 3 = average, 4 = poor, and 5 = nondiagnostic). Statistical comparisons between 3T, real 64mT, and simulated 64mT imaging were performed using a paired sample Wilcoxon signed-rank test (298).
Figure 20 - Validation of image transformation method. (A) Transformation applied to three novel hydrocephalus patients not used to develop the domain transformation. 3T images (column 1), 64mT images (column 2), and simulated 64mT images (column 3). (B) Transformation was applied to ten novel MS patients, three of which are visualized here. No patients visualized were included in the training set used to develop the automated transformation.

5.2.5 Modulating lesion contrast

Differences in magnetic relaxation times and optimal sequence parameters at LF relative to HF can contribute to differences in tissue and lesion contrasts (286). Although our domain transformation approach does not mathematically model these potential differences, we can use lesion segmentations to evaluate how changes in lesion contrast and conspicuity should be expected to affect detection accuracy. Thus, we prepared additional simulated LF images from the HGG dataset with decreasing signal intensity within the segmented lesions. We chose to run this sub-analysis on the BraTS HGG dataset because the glioma segmentations contained 4 tissue type labels, which allowed us to modulate contrast in each area separately, providing overall better
contrast modulation. Lesion intensity values were scaled independently from surrounding brain tissue in 20% increments over a range from 100% (original intensity) to 0% (isointense with background tissue). Isointensity was defined as mean lesion intensity equal to mean intensity of non-lesional tissue in the same slice.

Separate classifiers were trained to identify lesions at each intensity level, allowing for the decoupling of intensity contrast and structural abnormalities in classifier performance. Note that because of structural abnormalities caused by large tumors, such as mass effect, midline shift, and ventricular effacement, as well as intensity heterogeneity within lesions, even isointense lesions may retain some structural and signal alterations after contrast modulation. To minimize the effect of within lesion intensity heterogeneity, we restricted our analysis to patients within the top 50% of lesion homogeneity (N=130), as defined by within tumor signal to noise ratio (SNR). Accurate detection of isointense lesions should therefore be primarily driven by structural abnormalities.

While tissue relaxation rates of pathology remain unknown on the low-field system, this contrast modulation approach can be used to gauge how much contrast would be necessary for lesion detection. Importantly, while we applied this intensity modulation approach to alter contrast between pathology and background tissue, the same approach could be applied to gray matter, white matter, and cerebral spinal fluid segmentations to vary contrast between tissues.

5.2.6 Model Architecture

A convolutional neural network model was used to identify pathology in each high-field and simulated low-field dataset. Model construction and training was carried out using the Keras API (299) with TensorFlow (36,300) backend. Model architecture consisted of the DenseNet-121 network (301,302) with initial weights pre-trained on the ImageNet database (38) and four
additional densely-connected layers using Xavier initialization. This architecture was consistent across datasets.

### 5.2.7 Model Training

For each dataset, a unique model was trained to perform binary classification of axial slices (lesion present vs. lesion absent). To obtain the best estimate of device sensitivity as well as avoid scanner and site confounds, separate models were trained for each dataset rather than combining datasets and using a multi-class classifier. Separate models were trained on the HF and simulated LF images. Slices were labeled as “lesion present” if at least one pixel from the ground-truth lesion segmentations was present. Each dataset was divided (~9:1 split) into “training” and “test” datasets. Each patient was confined to either the training or test dataset. Training image order was randomly shuffled. All reported performance metrics are derived from held-out test data. Models were trained for 100 epochs using the Nadam optimizer, a learning rate of 0.002 with decay (303), and a batch size of 32. Batch size was chosen to accommodate VRAM of a Titan X GPU. Training data were augmented using random horizontal flipping. Training hyperparameters were consistent across all models.

### 5.2.8 Model Evaluation

Classification performance was evaluated using two metrics: (1) area under curve (AUC) of the receiver operating characteristic (ROC) and (2) F1 score (harmonic mean of precision and recall). A random chance null model (performance averaged over 1000 trials) was included for comparison. ROC curves were compared using DeLong’s test, implemented using the pROC R package (304). Logistic regression was used to quantify the impact of lesion size and intensity on detection. Significance of logistic regression parameters was determined by the Wald test.
Class activation maps (CAMs) were generated from shallow and deep convolutional layers to identify discriminative image regions (305). CAMs were constructed using the output feature map of a given convolutional layer with each feature map channel weighted by the lesion class gradient. Conceptually, CAMs help to interpret model function by visualizing image areas that are driving the model’s classification decision.

In addition to slice-by-slice classification performance, models were evaluated on a per-patient basis. For each test patient, the model assigned a classification score to all axial slices. A sliding convolutional filter was used to determine the mean classification score over several adjacent slices (approximately 1.5 cm in the z-axis).

5.2.9 Code and data availability


5.3 Results

5.3.1 Image transformation: Quantitative validation

Quantitative validation of the image transformation method was performed using data from the three additional hydrocephalus patients not included during the transformation fitting step. For
each participant, a standard 3T FLAIR image was transformed into a simulated LF image for comparison against the authentic LF ground truth. Representative images from each participant are shown in Figure 20A.

Image quality was quantitatively assessed using the entropy of the MR image gradient. Lower gradient entropy indicates sharper features and correlates strongly with radiologists’ perception of image quality on MRI (297). Gradient entropy of HF images (mean ± standard deviation: 5.91±0.54) was significantly lower than both the real LF images (7.89±0.90) and simulated LF (7.42±0.73) images (t-test, p<0.0001). While there was a statistical difference between the gradient entropy of real and simulated LF images (t-test, p<0.05), the effect size was dramatically reduced compared to the original HF images (0.47 vs 1.98). While perceived quality was modestly higher in simulated LF images, there was substantial overlap of gradient entropy with real LF images, indicating similar image quality between simulated and real images.

5.3.2 Image transformation: Radiologist validation

Representative image transformations from MS patients are shown in Figure 20B. Perceptions of diagnostic quality for real and simulated 64mT images (Figure 21) were similar (mean ± standard deviation, 2.98±1.04 and 3.04±0.86 respectively) with no statistical difference detected between the ratings (paired sample Wilcoxon signed-rank test, p=0.32). Both real and simulated 64mT images were rated as having average diagnostic quality, which was significantly lower than clinical 3T imaging, which was rated as having excellent quality (1.27±0.47, paired sample Wilcoxon signed-rank test, p<0.0001). White matter lesions were detected at similar rates in real and simulated 64mT images (86.6% and 82.2% respectively), both lower than the baseline detection rate of 94.4% in 3T imaging. Confidence in rating the presence or absence of lesions was slightly lower for simulated 64mT (1.88±1.08) compared to real 64mT (1.59±0.96) images (paired
sample Wilcoxon signed-rank test, p<0.01). Taken together, these results indicate that radiologists found perceived quality and clinical utility to be similar between real and simulated 64mT imaging.

Figure 21 - Radiologist ratings of real and simulated images. (A) Radiologists identified lesions at a similar rate in real (86.6%) and simulated (82.2%) 64mT images, both lower than 3T imaging (94.4%). (B) Image quality was also rated as similar between real and simulated 64mT images (no significant difference, paired sample Wilcoxon signed-rank test, p=0.32). Both real and simulated 64mT images were rated as having significantly lower quality than 3T imaging (paired sample Wilcoxon signed-rank test, p<0.001). (C) Raters were however more confident in their ratings for real 64mT compared to simulated images (paired sample Wilcoxon signed-rank test, p<0.01). Confidence ratings of both real and simulated 64mT images were significantly lower than 3T imaging (paired sample Wilcoxon signed-rank test, p<0.0001).

5.3.3 Comparing pathology detection in standard and simulated low-field images

Deep learning models were trained to perform binary classification of images in each disease cohort using either HF or simulated LF MRI. ROC curves for each cohort are shown in Figure 22. Despite significant image degradation, per-slice classifier performance was similar for HF and simulated LF MRI across all pathologies (Table 5). As expected, accuracy on more subtle pathology (MS lesions) was lower than more prominent pathology for both HF and simulated LF MRI datasets. The performance achieved is comparable to previous benchmarks using deep learning for detection of brain masses (306,307), and MS lesions (308), and significantly exceeded null models in all cohorts tested.
Figure 22 – Pathology detection performance. Receiver operating characteristic (ROC) curves shown for binary classification of images with and without pathology. (A) High-grade glioma (N=259) 3T classifier used 35,883 training images and 3,797 testing images, while the simulated 64mT classifier used 8,334 training images and 882 testing images. (B) Low-grade glioma (N=76) 3T classifier used 10,230 training images and 1,085 testing images, while the simulated 64mT classifier used 2,376 training images and 252 testing images. (C) Ischemic stroke (N=28) 3T classifier used 2,553 training images and 614 testing images, while the simulated 64mT classifier used 865 training images and 143 testing images. (D) Multiple sclerosis (N=20) 3T classifier used 10,824 training images and 1,428 testing images, while the simulated 64mT classifier used 812 training images and 124 testing images. No significant differences between 3T and simulated 64mT ROC curves were detected for any pathology. Abbreviations: True positive rate (TPR), False positive rate (FPR), Simulated (Sim.).

<table>
<thead>
<tr>
<th>Pathology</th>
<th>Standard AUC</th>
<th>Low-Field AUC</th>
<th>pAUC</th>
<th>Standard F1</th>
<th>Low-Field F1</th>
<th>Null Model F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>HGG</td>
<td>0.972</td>
<td>0.978</td>
<td>0.16</td>
<td>0.920</td>
<td>0.896</td>
<td>0.481±0.01</td>
</tr>
<tr>
<td>LGG</td>
<td>0.957</td>
<td>0.949</td>
<td>0.61</td>
<td>0.885</td>
<td>0.880</td>
<td>0.468±0.02</td>
</tr>
<tr>
<td>Stroke</td>
<td>0.936</td>
<td>0.940</td>
<td>0.83</td>
<td>0.772</td>
<td>0.761</td>
<td>0.485±0.02</td>
</tr>
<tr>
<td>MS</td>
<td>0.896</td>
<td>0.873</td>
<td>0.49</td>
<td>0.766</td>
<td>0.745</td>
<td>0.442±0.02</td>
</tr>
</tbody>
</table>

Table 5 – Performance metrics for each pathology type using standard or simulated low-field images. Abbreviations: High Grade Glioma (HGG), Low Grade Glioma (LGG), Multiple Sclerosis (MS), Area Under the Curve (AUC).
Figure 23 – Detection sensitivity as a function of lesion size and scaled intensity. Sensitivity of the deep learning classifiers for detecting lesions in the test set is shown in the (A) simulated LF (1401 testing slices) and (B) standard HF (6924 testing slices) images. Areas highlighting discrepancies between the datasets are highlighted in image insets. (C) Sensitivity of lesion detection in simulated LF images relative to HF images. Each point represents the sensitivity ratio measured on all lesions smaller than the given size threshold. Note that sensitivity is similar between image types when averaged over all lesions but differs significantly when restricted to smaller lesions.

5.3.4 Characterizing low-field pathology detection

To broadly characterize pathology detection capabilities in simulated LF images, detection sensitivity was aggregated across all pathologies and modeled using a logistic regression as a function of lesion size and intensity as shown in Figure 23A & 23B. For both HF and simulated LF images, sensitivity was more strongly associated with lesion size, though both parameters
reached statistical significance (standard: $z_{\text{size}}=13.5$, $p_{\text{size}}<2\text{e-16}$, $z_{\text{intensity}}=9.2$, $p_{\text{intensity}}<2\text{e-16}$; simulated low-field: $z_{\text{size}}=8.5$, $p_{\text{size}}<2\text{e-16}$, $z_{\text{intensity}}=3.9$, $p_{\text{intensity}}<9\text{e-5}$). HF imaging outperformed simulated LF imaging for detection of smaller or less intense lesions as shown in Figure 23C. While performance did not vary significantly between HF and simulated LF images across the cohorts as a whole, sensitivity differences in this subgroup analysis suggests a performance drop-off when using LF imaging for 1-4 cm² lesions (Figure 23A & 23B). These findings are agnostic to pathology type and may serve as generalizable performance thresholds for FLAIR at 64mT in yet-untested patient populations.

5.3.5 Patient-level classification

In addition to per-slice performance, we assessed pathology detection on a per-patient basis. Algorithms were evaluated over 17 held-out patients (three MS, four stroke, five HGG, five LGG) and achieved 100% sensitivity in both HF and simulated LF images. The LGG classifier was also evaluated using five control subjects, and correctly identified all subjects as non-lesional (100% specificity in both HF and simulated LF images) as shown in Figure 24A & 24B.

5.3.6 Class activation mapping

We used class activation mapping to probe which image regions were driving algorithm decisions as shown in Figure 24C. As expected, areas containing pathology are the primary drivers of classification at both shallow and deep layers. The deep CAM for the MS model also reveals that this model attends to periventricular areas known to be clinically important for lesion identification. These findings are reassuring that the tested models are detecting pathology of interest as expected and may have empirically captured features of the typical disease distribution. It is important to note that CAMs serve as an approximate representation of model attention, and each convolutional model layer has a unique CAM. While interpretation of CAMs alone is difficult
due to the nonlinear nature of neural networks, the CAMs and the patient level visualizations provide convergent evidence that our models are attending to pathological features.

**Figure 24 – Model validation and interpretability.** Panel A and B provide examples of per-patient pathology detection. Convolutional filters were used to generate average lesion probability values across several adjacent axial slices. A threshold value for patient-level classification was determined empirically by maximizing per-patient classification accuracy in the training set. Sample plots are shown for a patient with HGG (A) and a control patient (B) using simulated low-field imaging. (C) Class activation mapping. Row 1: Sample images of high-grade glioma (low-field), multiple sclerosis (low-field), low-grade glioma (standard), and ischemic stroke (standard). Row 2: CAMs generated from shallow network layer for each pathology. Row 3: CAMs generated from deepest convolutional network layer for each pathology.

5.3.7 Determining the effect of lesion intensity on detection

Again, a potential limitation of the simulated image generation method is that it does not account for possible changes in LF lesion to background tissue contrast relationships due to differences in relaxation times or pulse sequences. Here, we quantify detection robustness by measuring performance over a range of lesion contrasts (outlined in 5.2.5 Modulating lesion contrast) for HGG images. This approach also allows us to assess the relative importance of lesion to background tissue contrast in comparison to structural distortion from large brain tumors.
ROC curves for detection of HGGs are shown in Figure 25. Compared to detection of full-intensity tumors (AUC=0.972), there was a statistically significant decrease in performance for tumors with relative contrast of 60% or less (AUC_{60}=0.972, p=NS; AUC_{40}=0.943, p=0.01; AUC_{20}=0.905, p=2.0e-5; AUC_{0}=0.901, p=2.2e-6; AUC_{0}=0.874, p=2.0e-8). While contrast has a substantial impact on lesion detection, an AUC of 0.874 and F1 of 0.71 is achieved even for isointense lesions, indicating that in this dataset even with reduced lesion contrast large pathology could be identified due to structural deformation.

Figure 25 – Intensity modulation to explore effects of intensity contrast. (A) Intensity values of the pathology segmentation were modulated over a range from normal intensity (100%) to isointense (tumor = background). (B) For each image subset, a classifier was trained to distinguish pathological and normal slices (N=259, 8,334 training slices, 882 testing slices). AUC varied directly with lesion contrast but remained significantly better than chance even in the isointense cohort, which likely reflects structural deformations (such as ventricular effacement and midline shift as in the bottom row of panel A) or residual signal intensity heterogeneity.

5.4 Discussion

In this study we propose a generalizable method of image simulation and interpretation for the assessment of novel imaging modalities, particularly those that involve tradeoffs and lower image quality compared to an accepted standard, applied here to portable LF MRI. By leveraging open-access datasets, this virtual trial paradigm permits rapid, low-cost assessment of a device’s
potential diagnostic capabilities across a range of pathologies with limited real device data. We assert that this approach can help to address challenges in medical device development, regulatory approval, and clinical trial design.

Portable LF MRI offers an exciting opportunity for improving imaging accessibility in low-resource environments and enables point-of-care MRI. These scanners have relatively low manufacturing and operating costs and may help stem the increasing contribution of medical imaging to healthcare expenditures (309). To accelerate device development cycles and reduce the cost of bringing devices to market, it is pivotal that tools are developed to allow rapid prototyping, efficient regulatory approval, and expedited deployment. Virtual clinical trials can efficiently guide medical imaging technology evaluation by simulating patients, imaging systems, and image interpreters (310). For example, in breast tomography, Barufaldi et al. developed a virtual breast phantom and analytical pipeline that can simulate a clinical trial for several hundred patients per day (311). While not a replacement for traditional prospective studies, simulated clinical trials may offer significant value as the FDA considers devices prior to extensive prospective data collection. Simulated trials could contribute to early feasibility studies (EFS) or provide supporting evidence for investigational device exemption (IDE) approval. Additionally, simulated trials could identify key patient populations or indications to prioritize for standard clinical trials involving imaging. When considering a particular disease process, a simulated trial approach could help to establish benchmarks that a proposed device (such as a scanner or pulse sequence) must meet to provide sufficient diagnostic performance.

Based on simulated LF images, our study suggests LF MRI scanners should detect many brain lesions with comparable performance to standard MR imaging. However, simulated LF imaging was sensitive to lesion size. Accuracy was lower for 1-4 cm$^2$ lesions as shown in Figure 23A. These findings indicate that LF MRI may perform adequately for identifying macro-scale
pathology (most gliomas, medium-large vessel stroke, etc.) or measuring major brain structures, but may be less reliable for more subtle pathologies (small MS lesions, embolic infarcts).

We expect portable LF MRI to be used predominantly for clinical applications where standard MRI is either not feasible or delayed. For this reason and as a proof of concept, our analysis is limited to basic diagnostic capability (i.e. lesion sensitivity), defining a range of expected size and signal intensity thresholds, rather than more complex image interpretation such as distinguishing among pathologies, precise lesion segmentation, or tracking lesion evolution over time. While we compare performance of LF MRI to 3T MR devices, the practical alternative in certain use cases (ICU patients, underserved communities, in-office disease tracking) would be portable CT scans or no imaging at all. In these settings, LF MRI may have advantages over CT such as increased tissue contrast and lack of ionizing radiation exposure.

This work underscores the power of open-access clinical databases to facilitate translational research. Platforms for data sharing, such as XNAT Central (294), iEEG Portal (312), and crowdsourced competitions (313) have led to rapid advances in machine learning. While public databases provide diverse repositories of patient data with sufficient sample sizes to train deep learning algorithms, most medical imaging data remains federated across institutions (314). Further data-sharing efforts designed explicitly for evaluating devices and software for regulatory approval could reduce the cost and time necessary to bring innovative imaging technology to the clinic. Recently radiology has shifted toward centralizing algorithms while maintaining individual data ownership (315). While this approach may facilitate algorithm validation across research groups, it precludes creative use of multi-institutional data for applications beyond algorithm testing.

The simulated trial paradigm presented here is meant to serve as a framework for applying pre-existing datasets and deep learning to explore the expected performance of novel diagnostic
devices. However, the approach brings with it several important limitations and methodological considerations. Most importantly, the utility of simulated data is directly linked to the transformation method quality. Here, we implemented a relatively simple histogram matching based algorithm. While the present method approximates SNR and resolution in LF images, it does not account for other potential field strength, pulse sequence and device-specific artifacts that may affect lesion conspicuity and image quality.

T1 and T2* relaxation times vary as a function of magnetic field strength, which impacts intensity and contrast between tissues at low-field (286). Additionally, eddy currents and permanent magnet imperfections cause device specific artifacts. Neither tissue contrast differences nor image artifacts are modeled in our current approach. More advanced transformation methods, including generative adversarial networks (GANs), synthetic MRI, and other quantitative methods (316–319) could improve simulation quality and potentially overcome some of these limitations of our present approach.

However, transformation algorithms that learn by example, such as GANs, require large amounts of data. Methods that simulate images with a low N can be advantageous in certain situations (318). Specifically, low-data requirement methods can be useful for evaluating new devices or during the prototyping process, where available data are scarce. While data-driven methods may produce closer matched simulations in the long run, these methods may not be feasible for all applications. Importantly, the simulated trial approach is not meant to serve as a replacement for prospective clinical trials. While more advanced methods such as GANs may improve image-to-image simulation quality, there will still likely be a gap between real and simulated images, especially in patients with pathology (319). Simulations can provide useful guidance, including expected outcomes for prospective trials, however these retrospective analyses cannot provide the same level of scientific evidence as prospective clinical trials.
The scope of the present work is also limited to a specific field strength and pulse sequence. We only applied the domain transformation to 3T FLAIR imaging with the goal of simulating 64mT data. Image quality at low-field varies significantly depending on the pulse sequence, which means domain transformation results are unlikely to generalize across sequences. For instance, DWI is particularly susceptible to artifacts due to distortions, eddy current, and system stability. While more complex simulations incorporating multiple pulse sequences could be developed using the current framework, domain transformations would need to be developed for each sequence independently and it may be difficult to accurately simulate sequences that are significantly impacted by image artifacts. Additionally, our study only evaluated the domain transformation approach for 3T to 64mT. To validate that the approach is more broadly applicable, it should be evaluated across a range of field strengths such as 7T to 3T or 3T to 1.5T (320).

Additionally, special consideration should be taken when interpreting the machine learning results that compare 3T and simulated 64mT datasets. While performance of pathology detection was similar between datasets, pathology detection is only one component of diagnostic imaging, and these results should not be interpreted as equivalent clinical utility between the devices. Furthermore, our approach is limited to a single sequence and performs a simple detection task using separate classifiers for each pathology. Radiologists use multiple contrasts and incorporate patient information when making a clinical diagnosis. Additionally, while our approach likely provides the best estimate of device sensitivity and avoids potential scanner and site confounds associated with combining publicly available datasets, it does not provide insight into device specificity for distinguishing different pathologies. In future work, it may be possible to reconcile scanner and site differences using data harmonization methods (321–323).
5.5 Conclusion

In this study, we have proposed a method for guiding evaluation of imaging devices via simulated trials, incorporating domain transfer and automated pathology detection, and demonstrated its application to a new portable LF MRI device. This method allows for rapid evaluation of actual or proposed diagnostic imaging devices and can provide guidance and justification for prospective studies. In our simulations, we found that gliomas, strokes, and multiple sclerosis lesions could be detected in LF quality images and characterized size and signal intensity differences affecting lesion detection. This work additionally highlights the importance of centralized data sharing for device design and validation.
Chapter 6 – Portable, Low-Field MRI and Multiple Sclerosis

The recent development of portable, low-field MRI has potential to impact the management and screening of multiple sclerosis by changing how, where, and when patients are imaged. However, the sensitivity of portable, low-field MRI for multiple sclerosis lesions remains unknown. In this chapter, we use manual measurements and automated segmentation to evaluate the ability of a portable 64mT MRI scanner to characterize MS lesion burden and demonstrated effects of size and intensity on detection. This work appears as a preprint on medRxiv (171) and was published in *NeuroImage: Clinical* (324), with T. Campbell Arnold as first author.

**Portable, Low-Field Magnetic Resonance Imaging Sensitively Detects and Accurately Quantifies Multiple Sclerosis Lesions**

T. Campbell Arnold¹,², Danni Tu³, Serhat V. Okar⁴, Govind Nair⁴, Samantha By⁵, Karan Kawatra⁴, Timothy E. Robert-Fitzgerald³, Lisa M. Desiderio⁶, Matthew K. Schindler⁷, Russell T. Shinohara²,³, Daniel S. Reich⁴, Joel M. Stein²,⁶

1. Department of Bioengineering, University of Pennsylvania
2. Center for Neuroengineering and Therapeutics, University of Pennsylvania
3. Department of Biostatistics, University of Pennsylvania
4. Hyperfine, Guilford, CT
5. National Institute of Neurological Disorders and Stroke, National Institutes of Health
6. Department of Radiology, University of Pennsylvania
7. Department of Neurology, University of Pennsylvania
Abstract

Magnetic resonance imaging is a fundamental tool in the diagnosis and management of neurological diseases such as multiple sclerosis (MS). New portable, low-field MRI scanners could potentially lower financial and technical barriers to neuroimaging and reach underserved or disabled populations. However, the sensitivity of low-field MRI for MS lesions is unknown. We sought to determine if white matter lesions can be detected on a 64mT low-field MRI, compare automated lesion segmentations and total lesion burden between paired 3T and 64mT scans, and identify features that contribute to lesion detection accuracy. In this prospective, cross-sectional study, same-day brain MRI (FLAIR, T1, and T2) scans were collected from 36 adults (32 women; mean age, 50 ± 14 years) with known or suspected MS using 3T (Siemens) and 64mT (Hyperfine) scanners at two centers. Images were reviewed by neuroradiologists. MS lesions were measured manually and segmented using an automated algorithm. Statistical analyses assessed accuracy and variability of segmentations across scanners and systematic scanner biases in automated volumetric measurements. Lesions were identified on 64mT scans in 94% (31/33) of patients with confirmed MS. The smallest lesions manually detected were 5.7 ± 1.3 mm in maximum diameter at 64mT vs 2.1 ± 0.6 mm at 3T. Automated lesion burden estimates were highly correlated between 3T and 64mT scans (r = 0.89, p < 0.001). Bland-Altman analysis identified bias in 64mT segmentations (mean = 1.6 ml, standard error = 5.2 ml, limits of agreement = -19.0–15.9 ml), which over-estimated low lesion burden and under-estimated high burden (r = 0.74, p < 0.001). Visual inspection revealed over-segmentation was driven by flow-related hyperintensities in veins on 64mT FLAIR. Lesion size drove segmentation accuracy, with 93% of lesions >1.0 ml and all lesions >1.5 ml being detected. These results demonstrate that in established MS, a portable 64mT MRI scanner can identify white matter lesions, and disease burden estimates are consistent with 3T scans.
6.1 Introduction

Multiple sclerosis (MS) is a complex inflammatory and degenerative disease of the central nervous system (325). MS causes demyelinating lesions, typically assessed using magnetic resonance imaging (MRI). Imaging features related to white matter lesions (WML), such as number, volume, and dissemination in space and time, are key diagnostic criteria of MS (5) and determine treatment courses and clinical trial eligibility (6). Early diagnosis leads to better clinical outcomes, including delayed disease progression and reduced severity (326).

Although MS affects ~800,000 people in the United States (327) and likely >2.5 million people globally (328), the significant cost, infrastructure, and technical requirements associated with traditional high-field-strength MRI limit access to imaging worldwide (9). The scarcity is particularly felt in low-resource, sparsely populated, and rural areas (191). As the lack of diagnostic imaging can lead to delayed diagnosis and treatment, which result in worsening health disparities (12), there is renewed interest in low-field MRI, which employs magnets with a field strength of 1 tesla (T) or lower, as a lower-cost and potentially portable alternative to high-field MRI (136).

Recent improvements in hardware as well as image reconstruction and processing algorithms (122) have made low-field MRI promising in contexts where modest resolution is sufficient for diagnostic purposes (199). The clinical utility of portable low-field MRI has already been demonstrated for bedside monitoring in intensive care settings, where patients may not be stable enough to transport for traditional imaging (157,288,329). In the outpatient treatment of diseases such as MS, portable low-field MRI has the potential to lower barriers to accessing MRI technology and allow more frequent monitoring of disease activity (330), however its sensitivity and accuracy have not been explored. In particular, WML to background signal intensities and size thresholds for detection are unknown.
In this study, we assessed the feasibility of portable low-field MRI for MS lesion identification and lesion burden estimation. We collected paired same-day brain MRI from adults with MS at 3T and 64mT at two different institutions and compared lesion detection using both manual and automated measurements. We anticipated that MS lesions would be detectable at 64mT, though sensitivity to small lesions would likely be reduced. Finally, we explored a simple approach for super-resolution imaging of small lesions based on multi-acquisition image averaging.

### 6.2 Materials and Methods

#### 6.2.1 Participants & Imaging

Among adult outpatients undergoing clinical 3T brain MRI for known or suspected MS between October 2020 and April 2021, 36 patients (Figure 26) were recruited at site A (N=21) and site B (N=15). All patients received same-day 3T and 64mT MRI. Demographic information was collected from clinical notes and included age, sex, race, clinical phenotype, disease duration, Expanded Disability Status Scale (EDSS), and current disease modifying therapy (Table 6). This study was approved by each site’s institutional review board, and patients provided written, informed consent.
Figure 26 – Flow chart of study participants. Abbreviations: multiple sclerosis (MS), clinically isolated syndrome (CIS), neuromyelitis optica (NMO).

High-field MRI was performed on 3T scanners (Siemens, Erlangen, Germany). Each site used a standardized, whole-brain imaging protocol, which included 3D T1-weighted (T1w), T2-weighted (T2w), and 3D T2-FLAIR sequences (Figure 27A). Sequence parameters are listed in Table 7. At site B, patients received gadolinium (gadobutrol, 0.1 mmol/L) prior to 3T scanning; 64mT scans were obtained after contrast-enhanced 3T scans with mean post-gadolinium duration of 58 ± 21 minutes.

Same-day, low-field MRI was performed on portable 64mT Swoop MRI systems (Hyperfine, Guilford, CT). Whole-brain, 3D T1w, T2w, and T2-FLAIR scans, analogous to those
acquired at 3T, were collected (Figure 27B). Including localizer and pre-scan calibration, total scan time was 26:06 minutes.

**Figure 27 – MS lesions appear similar on 3T and 64mT pulse sequences.** Paired 3T (A) and 64mT (B) images from a 66-year-old woman with stable RRMS showing deep gray matter lesions (short arrow) and periventricular white matter lesions (long arrow) on both scanners. Sequences include T1-weighted (left), T2-weighted (center), and T2-FLAIR (right).
<table>
<thead>
<tr>
<th></th>
<th>Total (N = 36)</th>
<th>Site A (N = 21)</th>
<th>Site B (N = 15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>49.6 ± 14.2</td>
<td>45.3 ± 13.6 *</td>
<td>55.7 ± 12.7 *</td>
</tr>
<tr>
<td>Sex (women/men)</td>
<td>32/4</td>
<td>19/2</td>
<td>13/2</td>
</tr>
<tr>
<td>Race/ethnicity (White/Black/Hispanic)</td>
<td>26/9/1</td>
<td>14/7/0</td>
<td>12/2/1</td>
</tr>
<tr>
<td>Disease duration (years)</td>
<td>13.7 ± 11.2</td>
<td>10.2 ± 9.6 *</td>
<td>18.5 ± 11.5 *</td>
</tr>
<tr>
<td>EDSS (0–10)</td>
<td>1.5 (IQR = 2.0)</td>
<td>1.5 (IQR = 2.0)</td>
<td>2.0 (IQR = 1.25)</td>
</tr>
<tr>
<td>Phenotype</td>
<td>RRMS (18), CIS (1), NMO (1), RIS (1)</td>
<td>RRMS (10), SPMS (2), CIS (1), ITM (1), PPMS (1)</td>
<td></td>
</tr>
<tr>
<td>Current disease modifying therapy</td>
<td>ocrelizumab (9), natalizumab (2), other (6), none (4)</td>
<td>dimethyl fumarate (6), ocrelizumab (3), other (3), none (6)</td>
<td></td>
</tr>
</tbody>
</table>

**Table 6 – Patient Demographics.** Demographic information and clinical history for 36 consecutive MS patients included in the study. An asterisk indicates a significant difference between sites. Abbreviations: Expanded disability status scale (EDSS), relapsing-remitting multiple sclerosis (RRMS), primary progressive multiple sclerosis (PPMS), and secondary progressive multiple sclerosis (SPMS), clinically isolated syndrome (CIS), radiologically isolated syndrome (RIS), neuromyelitis optica (NMO), idiopathic transverse myelitis (ITM).
<table>
<thead>
<tr>
<th>Sequence</th>
<th>Site</th>
<th>Field Strength (T)</th>
<th>TE (ms)</th>
<th>TR (s)</th>
<th>TI (s)</th>
<th>Resolution (mm)</th>
<th>Scan-time (min:sec)</th>
<th>Averages</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1w-MPRAGE</td>
<td>A</td>
<td>3</td>
<td>2.48</td>
<td>1.9</td>
<td>0.9</td>
<td>1x1x1</td>
<td>4:18</td>
<td>1</td>
</tr>
<tr>
<td>T1w-MP2RAGE</td>
<td>B</td>
<td>3</td>
<td>2.92</td>
<td>5</td>
<td>0.7, 2.5</td>
<td>1x1x1</td>
<td>8:30</td>
<td>1</td>
</tr>
<tr>
<td>T1w Both</td>
<td></td>
<td>0.064</td>
<td>6.26</td>
<td>1.5</td>
<td>0.3</td>
<td>1.5x1.5x5</td>
<td>4:52</td>
<td>1</td>
</tr>
<tr>
<td>T2-FLAIR 3D</td>
<td>A</td>
<td>3</td>
<td>398</td>
<td>5</td>
<td>1.6</td>
<td>1x1x1</td>
<td>5:02</td>
<td>1</td>
</tr>
<tr>
<td>T2-FLAIR 3D</td>
<td>B</td>
<td>3</td>
<td>352</td>
<td>4.8</td>
<td>1.8</td>
<td>1x1x1</td>
<td>7:15</td>
<td>1</td>
</tr>
<tr>
<td>T2-FLAIR Both</td>
<td></td>
<td>0.064</td>
<td>200</td>
<td>4</td>
<td>1.4</td>
<td>1.6x1.6x5</td>
<td>9:29</td>
<td>1</td>
</tr>
<tr>
<td>T2w A</td>
<td></td>
<td>3</td>
<td>103</td>
<td>5.5</td>
<td>N/A</td>
<td>0.5x0.5x5.2</td>
<td>1:46</td>
<td>2</td>
</tr>
<tr>
<td>T2w B</td>
<td></td>
<td>3</td>
<td>82</td>
<td>5</td>
<td>N/A</td>
<td>0.34x0.34x3</td>
<td>4:30</td>
<td>1</td>
</tr>
<tr>
<td>T2w Both</td>
<td></td>
<td>0.064</td>
<td>209</td>
<td>2</td>
<td>N/A</td>
<td>1.5x1.5x5</td>
<td>7:01</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table 7 – Sequence parameters for study scans.** Abbreviations: Tesla (T), T1-weighted (T1w), T2-weighted (T2w), Fluid-attenuated inversion recovery (FLAIR), echo time (TE), repetition time (TR), inversion time (TI).

6.2.2 Manual Review and Lesion Measurement

MRI scans were reviewed for WML by two neuroradiologists (DSR and JMS, 19 and 8 years experience). Maximum diameters (Dmax) of the smallest and largest WML visually detectable at each field strength were manually measured by a neuroradiologist (JMS) and a neurologist (SVO) with MS MRI expertise (3 years experience) using ITK-SNAP (35). All measurements were made on T2-FLAIR scans. Lesions were assessed in axial planes as well as sagittal and coronal reformations, and Dmax measurements were made on the plane with the largest lesion diameter. In confluent periventricular lesions, Dmax was measured perpendicular to the ventricle. Low-field imaging was evaluated prior to 3T scans to avoid interpretation bias, and image
sets were reviewed separately. Inter-rater reliability was assessed using two-way random, single-measure intraclass correlation coefficients (ICC) with 95% confidence intervals (CI) reported. Patients that did not meet the McDonald criteria for dissemination of lesions in space were excluded from subsequent analyses (5).

6.2.3 Automated Lesion Segmentation

The same segmentation pipeline was applied to 3T and 64mT images. Images were preprocessed using N4 bias correction (331), and each T2-FLAIR volume was rigidly registered to the corresponding T1w volume using Advanced Normalization Tools (ANTs) (32,45). A brain mask was obtained using Multi-Atlas Skull-Stripping (MASS) (332). Finally, to enable comparisons across patients, image intensities were normalized using White Stripe (333) within each sequence. Lesion segmentation was performed using the Method for Inter-Modal Segmentation Analysis (MIMoSA) (334,335), an automated pipeline that leverages shared information (coupling) between modalities to produce probability maps of WML (Figure 28). All probability maps were thresholded at 0.2 to generate binary lesion masks, manually selected based on prior empirical evidence.

![Figure 28 - Automated segmentation algorithm. Visual description of the steps in the MIMoSA algorithm.](image)
6.2.4 Automated Segmentation Evaluation

Estimation of total lesion burden was the primary performance measure compared between 3T and 64mT. Two lesion burden estimates were obtained for each patient by calculating lesion segmentation volumes for the respective scanners. The relationship between volume estimates was assessed using Pearson’s correlation. Bland-Altman plots were used to determine agreement and assess for systematic scanner biases.

Similarity between segmentation masks was assessed using the Dice-Sørensen coefficient (Dice), which measures the overlap between two images ($X$ and $Y$):

$$Dice = \frac{2|X \cap Y|}{|X| + |Y|}. \quad (1)$$

Dice scores range from 0–1, with 1 indicating perfect segmentation overlap. Prior to Dice calculation, 3T and 64mT images were coregistered using ANTs (45). While Dice score may not reflect segmentation quality when the number of target objects is not known a priori, this measure has been included as it is widely used and allows for comparisons across studies (336). All 3T and 64mT segmentations were manually reviewed to verify overlapping regions were WMLs rather than false positive detections.

6.2.5 Size and Intensity Analysis

Connected-components analysis was used to identify individual lesions in automated 3T and 64mT segmentations (337). Sensitivity to individual lesions at low-field MRI was assessed using the true-positive rate (TPR), or the proportion of lesions correctly identified:

$$TPR = \frac{TP}{TP + FN}. \quad (2)$$
where true positives \((TP)\) are defined as lesions where 64mT and 3T segmentations overlap and false negatives \((FN)\) are defined as lesions with 3T segmentation but no 64mT segmentation overlap. The false-discovery rate (FDR) was assessed as:

\[
FDR = \frac{FP}{FP + TP},
\]

where false positives \((FP)\) are defined as lesions with 64mT segmentation but no 3T segmentation overlap. Lesion overlap was defined as at least one shared voxel between the 3T and 64mT lesion segmentations. To understand the impact of lesion features on detection rates, TPR and FDR were plotted as a function of lesion size and normalized lesion intensity (333).

### 6.2.6 Super Resolution Imaging

Low-field MRI necessitates a trade-off between signal-to-noise ratio and image resolution, which limits the minimum detectable lesion size. However, image quality can be increased by taking advantage of partial volume effects in multiple scans (338). In one patient, multi-acquisition volume averaging was explored. Eight 64mT T2-FLAIR acquisitions (TE = 1.8 s, TR = 4 s, TI = 1.4 s, averages = 80, scan time = 6:03 min, resolution = 1.8x1.8x5 mm) were collected with head repositioning between scans (total scan time: 48:24 minutes). Images were resliced to 1 mm isotropic resolution with linear interpolation, underwent affine registration to the initial acquisition, and were averaged to create a single high-resolution image. Super resolution images were iteratively generated for each additional acquisition, for a total of eight images.

To quantify lesion conspicuity, we manually segmented the white matter lesion and a similarly sized region in normal appearing ipsilateral white matter on 3T imaging using ITK-SNAP (35). The 3T image and lesion segmentations were registered to the initial 64mT acquisition. We calculated lesion conspicuity as the ratio of the difference and the sum of mean intensity in the two segmentations:
\[ Conspicuity = \frac{X - Y}{X + Y} \]  

where \( X \) is the mean intensity in the lesion segmentation and \( Y \) is the mean intensity of the normal appearing white matter segmentation. The segmentations were applied to each super resolution iteration to calculate conspicuity for that iteration.

### 6.2.7 Statistics and Data/Code Availability

All code related to this study is publicly available. The MIMoSA algorithm is available in R as a Neuroconductor package ([https://neuroconductor.org/package/details/mimosa](https://neuroconductor.org/package/details/mimosa)) and on GitHub ([https://github.com/avalcarcel9/mimosa/](https://github.com/avalcarcel9/mimosa/)). T-tests, Pearson’s correlation, and summary statistics were calculated using scipy (v1.5.2) and numpy (v1.19.2) in Python (v3.8.5). Bland-Altman plots were visualized using pyCompare (v1.5.1) while boxplots and correlations utilized seaborn (v0.11.0). Inter-rater reliability was calculated using irr (v0.84.1) in R (v4.0.3). A manuscript companion containing all analyses is available on GitHub ([https://github.com/penn-cnt/Arnold_LF-MRI_MS](https://github.com/penn-cnt/Arnold_LF-MRI_MS)). The data generated in this study can be made available, with protected health information removed, upon reasonable request to the corresponding author and with a data sharing agreement between each institution in place.

### 6.3 Results

#### 6.3.1 Patient Demographics

We collected data from 36 adults with known or suspected MS. The patient population had a mean age of 49.6 (SD: 14.2) years and was composed of 32 women and 4 men (Table 1). The mean duration of disease was 13.7 years (SD: 11.2), and patients had a median EDSS of 1.5 (interquartile range = 2). Patients from site B were significantly older than those from site A (two-
sample t-test, \( t=2.2, \ p = 0.03 \), site A: 45.3 years old, site B: 55.7 years old) and had a correspondingly longer duration of disease (two-sample t-test, \( t=2.3, \ p = 0.03 \), site A: 10.2 years, site B: 18.5 years). Additional demographic information is provided in Table 1. After initial scan review, three patients were excluded from further analysis: One patient had excessive motion in the scanner and two patients did not meet the MS diagnostic criteria of having lesion dissemination in space (DIS) (5). All subsequent analyses are based on the remaining 33 patients (Figure 26).

6.3.2 Manual Measurements

MS lesions on 64mT are characterized by T1w hypointensity and T2w/T2-FLAIR hyperintensity, similar to 3T imaging (Figure 27). At 64mT, lesions were identified by at least one rater in 94\% (31/33) of patients with confirmed lesions on 3T imaging. In one patient, only one rater identified lesions at 64mT; all other low-field ratings were concordant. The largest and smallest lesions in each image were identified, and the Dmax was recorded. The 64mT scanner showed 100\% sensitivity for detecting WML when there was at least one lesion with Dmax >5 mm (31/33 patients, 94\%). Across patients, there was no significant difference in Dmax for the largest lesions measured at 64mT (15.1 ± 5.9 mm) and 3T (14.8 ± 6.6 mm) (Figure 29A). However, the mean Dmax for the smallest WML detectable was significantly larger (paired t-test, \( t=19.6, \ p < 0.001 \)) on 64mT (5.7 ± 1.3 mm) compared to 3T (2.1 ± 0.6 mm) (Figure 29B). There was no effect of scan site on Dmax measurements, however there was a significant difference between rater 1 (2.3 ± 0.5 mm) and rater 2 (1.9 ± 0.6 mm) for the smallest lesion detected at 3T (paired t-test, \( t=4.8, \ p < 0.001 \)). No gadolinium enhancing lesions were seen on 3T or 64mT imaging.
Raters from each site independently measured the maximum diameter (Dmax) of the smallest lesion (Sm) and largest lesion (Lg) in 3T and 64mT imaging for all patients. (A) For the largest lesion measurements, there were no significant differences between raters at 3T (t = 1.3, p = 0.19) or 64mT (t = 1.2, p = 0.23); additionally, there was no difference between 3T and 64mT measurements (t = 0.04, p = 0.97). (B) For the smallest lesion measurements, there was a significant difference between raters for 3T measurements (t = 4.83, p < 0.001) although 64mT measurements were not significantly different (t = 1.67, p = 0.11); additionally, the diameter of the smallest lesion was significantly lower (t = 19.6, p < 0.001) when measured on 3T (mean 2.1 mm) compared to 64mT (mean 5.7 mm). (C) Across all lesions there was a strong correlation (r = 0.90, p < 0.001) between raters. There was significant intraclass correlation for the largest lesion at 3T (ICC = 0.77, CI = [0.58-0.88]), largest lesion at 64mT (ICC = 0.91, CI = [0.83-0.96]), smallest lesion at 3T (ICC = 0.62, CI = [0.12-0.83]), and smallest lesion at 64mT (ICC = 0.66, CI = [0.4-0.82]), indicating a high degree of agreement between rater measurements for both 3T and 64mT.

6.3.3 Interrater Reliability

The smallest and largest lesion in each image were independently measured by two raters to assess interrater reliability (Figure 29). The ICC for each patient’s largest lesion measured at 3T and 64mT was 0.77 (CI = [0.58-0.88]) and 0.91 (CI = [0.83-0.96]) respectively, indicating high interrater reliability for large lesions on both scanners (Figure 29A). Similarly, when measuring each patient’s smallest lesion there was a significant relationship between raters at both 3T (ICC = 0.62, CI = [0.12-0.83]) and 64mT (ICC = 0.66, CI = [0.4-0.82]) (Figure 29B). This indicates that measurements made by raters had a similar degree of reliability at 3T and 64mT. Of note, the
average smallest lesions detected (3T: 2.1 ± 0.6 mm, 64mT: 5.7 ± 1.3 mm) approached the respective resolution limits for each system (3T: 1 mm, 64mT: 5 mm).

6.3.4 Total Lesion Burden Estimates

To obtain more objective measures, 3T and 64mT image sets were processed with the same automated lesion segmentation algorithm. Initial qualitative review of segmentation overlays revealed similar patterns of lesion segmentation, particularly with respect to large periventricular lesions (Figure 30). Quantitative comparisons indicated that estimates of total lesion burden were highly correlated (r = 0.89, p < 0.001) (Figure 31A). Mean lesion burden estimates were not significantly different (paired-t-test, t = 1.0, p = 0.32) between 3T (11.9 ± 16.5 ml) and 64mT (13.5 ± 10.2 ml) images.
**Figure 30 - Automated lesion segmentations at 3T and 64mT overlap.** (A) 64mT FLAIR images for three cases (left) with automated lesion segmentations generated from the 64mT images using MIMoSA overlaid (right). (B) Corresponding 3T FLAIR images for the same three cases (left) with 3T based segmentations (right). Patients from top to bottom are a 51-year-old female with RRMS, 44-year-old female with RRMS, and 71-year-old female with RRMS. All images were coregistered to 64mT T1-weighted images for comparison. Segmentations generated from 64mT and 3T scanners show similar patterns.

A Bland-Altman plot for agreement between 3T and 64mT lesion burden estimates is presented in **Figure 31B**. The mean difference was 1.6 ml with a 5.2 ml standard error of measurement, and the 95% limits of agreement were -19.0 to 15.9 ml. There was a significant correlation (r = 0.74, p < 0.001) between pairwise differences and averages, indicating that compared to 3T, the 64mT segmentations overestimate low lesion burdens and underestimate high lesion burdens. Visual inspection revealed that false positives contributing to over-segmentation were predominantly due to flow-related high signal intensity in veins, hyperintensity in non-lesion
structures (such as the pineal gland), and areas of artifactual peripheral high signal in cortical/subcortical tissue on 64mT FLAIR sequences (Figure 32).

Figure 31 - Total lesion volume measured at 3T and 64mT shows agreement. (A) Pearson’s correlation ($r = 0.89$, $p < 0.001$) and (B) Bland-Altman plot showing agreement between 3T and 64mT segmentation volumes (bias -1.6 ml, standard error of measurement = 5.2 ml, 95% limit of agreement -19.0 to 15.9 ml). Pearson’s correlation ($r = 0.74$, $p < 0.001$) in dark blue indicates over-segmentation at 64mT when lesion burden is low and under-segmentation when burden is high.
Figure 32 - 64mT image artifacts. One subject was excluded due to significant image artifacts at 64mT. The artifacts appeared bilaterally and were present on both FLAIR (left) and T1w (right) imaging. The T1w sequence was repeated several times and artifacts persisted throughout the acquisitions; after this the 64mT protocol for this patient was terminated. The patient was not wearing jewelry and did not have metal implants. We believe the artifacts may have been caused by a nearby large metal box, used to house the portable MRI when not in use.

6.3.5 Automated Segmentation Overlap

Across patients, there was a wide range in overlap between 3T and 64mT segmentation pairs (Dice: mean = 0.23, standard deviation = 0.21, max = 0.65, min = 0), with automated segmentations overlapping in 91% (30/33) of patients. Two patients had no segmentation overlap and one patient was excluded because the overlapping region was a hyperintense pineal gland, not a WML. All three patients without lesion overlap were in the bottom 12% of total lesion burden. To characterize the full range of segmentation quality across the dataset, Figure 33 illustrates segmentations from each quartile of the Dice distribution. Larger lesion size is frequently
associated with higher Dice (336). We found in our dataset that total lesion burden at 3T was highly correlated with Dice ($r = 0.81$, $p < 0.001$) (Figure 34).

**Figure 33 - False positive lesion segmentations on 64mT T2-FLAIR.** This figure illustrates common causes of 64mT false positive lesion detections on the top row with 3T comparisons provided on the bottom row. False positive detections were primarily caused by artifactual peripheral hyperintensity, as seen in (A) a 33-year-old female with RRMS, and hyperintense venous structures, as seen in (B) a 35-year-old female with RRMS, (C) a 55-year-old male with RRMS, and (D) a 50-year-old male with clinically isolated syndrome. Other hyperintense structures, such as a pineal cyst in panel D, were also a source of false positive labeling in both 3T and 64mT images.
Figure 34 - Dice score increases with lesion volume. Previous studies have found that larger lesion sizes are associated with higher Dice scores (336). In our study, we found a similar effect ($r = 0.81$, $p < 0.001$) such that subjects with higher lesion volume had correspondingly higher Dice scores.

6.3.6 Lesion Sensitivity and False Discovery

In each segmentation, individual lesions were identified using connected-components analysis (337). For each lesion, volume and mean intensity were quantified. The true-positive rate (TPR) and false-discovery rate (FDR) were calculated across a range of lesion size and intensity thresholds (Figure 35). The TPR increases dramatically with lesion size, reaching 93% for lesions >1 ml and 100% for lesions >1.5 ml. The FDR decreases with lesion size, reaching 36% for lesions >1 ml, 22% for lesions >1.5 ml, and 3% for lesions >2.5 ml. TPR also increases with mean lesion intensity, indicating that lesion intensity influences sensitivity; however, FDR remains high (>75%) regardless of the sensitivity, indicating a large number of false positive detections across intensity thresholds. Examples of false positive detections can be seen in Figure 33.
Figure 35 - Lesion size and intensity influence detection rate. (A) The detection rate, or true positive rate (TPR), steadily increases with lesion size, with 93% detected at >1 ml, and all lesions greater than 1.5 ml being detected. The false discovery rate (FDR) decreases with lesion size, with 36% false discovery rate at >1 ml, 22% at >1.5 ml, and 3% at >2.5 ml. Though the x-axis was limited to 4 ml for illustrative purposes, lesions >20 ml were found in the dataset. (B) To analyze the relationship between lesion intensity and detection rate, image intensity values were first normalized using White Stripe (Shinohara et al. 2014). While detection rate increases as mean lesion intensity increases, the FDR remains high (>75%) across lesion intensities. The high number of false positive detections was driven by hyperintense veins and peripheral signal artifacts, as seen in figure 33.

6.3.7 Super Resolution Imaging

In one patient, a 3x4x5 mm (0.06 ml) subcortical lesion was evident near the left middle frontal gyrus on 3T (Figure 36A) but not in any single low-field acquisition (Figure 36D). After multi-acquisition volume averaging of 3 to 8 acquisitions, the lesion became detectable on the low-field system, and lesion intensity relative to ipsilateral white matter steadily increased with additional acquisitions. With 8 volume averages, there was a 53% increase in lesion conspicuity, which was equivalent to 72% of conspicuity at 3T (Figure 36B). With multi-acquisition image averaging, lesions as small as 0.06 ml were discernible on 64mT scans.
Figure 36 - Multi acquisition image averaging increases resolution. This figure depicts a 3x4x5mm (0.06 ml) subcortical left frontal white matter lesion in a 53-year-old woman with stable RRMS and compares 64mT FLAIR images generated from multi-acquisition image averaging to 3T imaging. The lesion is readily apparent on 3T imaging (A); however, it could not be discerned in a single 64mT acquisition (D). Volume averaging of multiple acquisitions with repositioning between scans did reveal the lesion on the low-field system (B & C). The lesion was discernible for N≥3 multi acquisition averages. The lesion was manually segmented on 3T, and the ratio of mean lesion intensity to ipsilateral adjacent white matter (WM) is given as an estimate of lesion conspicuity (red dot). In 64mT imaging, the ratio steadily increases with additional acquisition averages (blue dots). With 8 volume averages, there was a 53% increase in lesion conspicuity.

6.4 Discussion

In this study, we compared manual and automated lesion detection in paired 3T and portable 64mT brain MRI from patients with known or suspected MS at two sites. On visual inspection of 64mT images, neuroradiologists were able to detect white matter lesions in 94% (31/33) of patients with discernable 3T lesions. An automated lesion segmentation algorithm detected overlapping lesions in 91% (30/33) of patients, and estimates of total lesion burden were highly correlated between 3T and 64mT scans (r = 0.89, p < 0.001). Together, these results suggest that portable 64mT imaging could have diagnostic utility in the context of MS.
Our investigation is motivated by recent advances in hardware development and reconstruction software that address the reduced signal-to-noise and resolution associated with low-field MRI. Commercial MRI systems have predominantly trended towards higher field strengths, with low-field systems relegated to niche applications (9,122,134). Currently, there is renewed commercial interest in developing low-field MRI systems (112), including Hyperfine’s portable 64mT Swoop system, Synaptive Medical’s 0.5T Evry system, and Siemens’s 0.55T Magnetom Free.Max system, all of which have received FDA clearance since 2020. As evidence indicates that low-field systems can detect relatively subtle pathologies, such as CNS demyelination (168,285,339), we were motivated to investigate the sensitivity of the newly available portable 64mT MRI for MS lesions.

Our study found that lesions could be identified on 64mT scans in 94% of patients with discernible lesions at 3T. Additionally, we found a strong correlation between total lesion burden estimates on 3T and 64mT scanners. However, the smallest detected lesion size was significantly larger at 64mT (5.7 ± 1.3 mm) compared to 3T (2.1 ± 0.6 mm), indicating that smaller lesions are missed at low field. Taken together, these findings suggest that the 64mT device may be useful for tracking disease severity over time, although the device may be less suitable for making an initial diagnosis when high-field scanners are available. However, while high-field MRI will likely remain the diagnostic tool of choice in population centers of developed countries, the lower costs and infrastructure requirements of portable low-field MRI could expand clinical options for patients in low and middle income countries and rural areas (12). Within the United States, 60% of rural hospitals lack an on-site MRI (107). Additionally, most MS patients will experience mobility impairment, which can impact quality of care (108). Mobile MRI units could bring imaging to patients, providing otherwise unavailable service to sparsely populated areas and individuals who cannot travel (181,201,340).
The lower cost and ease of use associated with low-field MRI could also impact how early and frequently patients are imaged (341). Rovira et al. were able to predict which individuals would develop clinically definite MS using an MRI collected less than 3 months after the patient’s initial symptom onset (342). Increased scan frequency could also potentially permit earlier assessment of therapy response or detection of treatment complications, such as progressive multifocal leukoencephalopathy (330). Further studies should assess low-field MRI sensitivity for new or growing lesions over time in each clinical scenario. Our work indicates an ability to detect individual lesions at least as small as 0.06 ml using super-resolution imaging.

Low-field MRI also offers the potential to conduct large-scale studies or screening of high-risk individuals at lower cost. In MS, high-risk asymptomatic family members have an increased incidence of neurological dysfunction and neuroimaging findings associated with MS (343). Additionally, patients with radiologically isolated syndrome (i.e., patients who meet MS criteria radiologically but are clinically asymptomatic) are known to be at high risk for development of clinical MS (344). However, studies of asymptomatic individuals require large sample sizes, which cause recruitment and cost restraints. The reduced cost of low-field MRI could significantly impact the type of population based and longitudinal studies available to researchers (341).

While machine learning methods for MS lesion segmentation have yet to consistently outperform manual segmentation, they reduce the cost, time, and subjectivity associated with manual labeling (334,335). Combining low-field MRI with automated techniques can further address barriers to MRI access and image interpretation. In our work, the average Dice overlap between automated 3T and 64mT segmentations was only 0.23, with three subjects having no overlap. The low overlap was driven in part by hyperintense venous structures and peripheral artifacts on 64mT FLAIR imaging, which also resulted in a higher number of false positives (22% for lesions >1.5ml) despite comparatively high lesion sensitivity (100% for lesion >1.5 ml). Our
work illustrates the importance of reassessing algorithm performance for low-field MRI sequences and indicates that retraining or tuning may be necessary to address differences in image quality and tissue contrast. In practice, an automated segmentation could serve as a biomarker for determining eligibility or endpoints in clinical trials or as a starting point for further manual refinements.

Whether gadolinium can be used to assess contrast-enhancing lesions on the 64mT device remains unknown. In our study, we saw no contrast-enhancing lesions on 3T or 64mT. Patients at site B received a standard dose of 0.1 mmol/L of a contrast agent (gadobutrol) optimized for high-field MRI; however, since 3T imaging was performed first, there was an average delay of 58 ± 21 minutes between contrast administration and 64mT imaging. At lower field strengths, the benefit from T1 shortening contrast agents is significantly reduced, and a higher gadolinium dosage (345), or potentially an alternative high-relaxivity agent (345,346), could be useful for low-field MRI. Additionally, the 58-minute delay approaches gadolinium’s half-life and this washout period can further attenuate the signal. Studies with minimal post-contrast delay and potentially higher doses of gadolinium could be conducted to better assess the potential for lesion enhancement at low field. However, even at higher doses, low-field devices may have reduced sensitivity to contrast-enhancing lesions (347).

The current study has several limitations. Our findings suggest that portable 64 mT FLAIR scans will be sensitive for white matter lesions in MS and more generally, but we focused on patients with established MS and did not assess the specificity of MS lesion detection relative to other disease processes or normal controls. In addition, sensitivity at the patient or lesion level will depend on the lesion burden and size distribution (291). We used automated 3T segmentation as ground truth, though complete labeling accuracy is challenging even at high field, and we considered lesion overlap and volume rather than
lesion counts. We only evaluated a single lesion segmentation method; results may not generalize to other algorithms. Indeed, our findings indicate that both image acquisition strategies and segmentation methods can be further optimized to increase the sensitivity and accuracy of low field lesion detection for larger prospective studies. We did not assess longitudinal imaging or the ability to detect new or active lesions. Gadolinium was only administered at one of the two sites and was not administered directly for 64mT imaging. As discussed above, given that none of the patients in our cohort had gadolinium enhancing lesions on their high-field scans and the post-contrast delay before each 64mT scan, we cannot assess whether contrast enhancing lesions can be seen at 64mT.

6.5 Conclusion

In conclusion, increased imaging capabilities of low-field MRI systems warrants their re-evaluation across a range of pathologies and indications. We found that a portable 64mT scanner was sensitive to WML and that an automated algorithm designed for 3T image segmentation can be applied to 64mT data. Although additional work will be needed to evaluate portable low-field MRI systems and their capacity to carry out specific clinical functions, our findings suggest promising avenues to more accessible imaging technologies for MS around the world.
Chapter 7 - Conclusion

Medical imaging remains underutilized and inaccessible for millions of patients throughout the world. A staggering billion people suffer from neurological disorders worldwide, and they need medical imaging to be accurately diagnosed and treated. Given the current cost and time investments necessary to obtain and evaluate MRI, we will not be able to scale this technology to meet the demands of society. In this work, I present methods for automated analysis of medical imaging and evaluate the clinical utility of a lower-cost portable MRI.

By taking advantage of recent advances in machine learning, we can automate tedious, labor-intensive tasks, thus allowing trained experts to increase imaging throughput and focus more on patient care. In Chapter 2, I developed a fully automated approach to segmenting the resection zone in postoperative epilepsy patients and determine what brain tissues were affected by the procedure. Integrating tools, such as this, into the clinical workflows and monitoring their impact on patient outcomes, healthcare costs, and physician experience will be a crucial next step. Neuroimaging has shown remarkable potential for diagnosing patients and providing evidence for the best course of treatment, as demonstrated in our work on epilepsy surgery in Chapter 3. However, these tools must be translated into clinical practice, either through commercial avenues or open-source development projects, in order to impact imaging efficiency and patient care.

The cost of MRI devices is another limiting factor to the expansion of medical imaging access in lower-resource areas. Most 1.5T and 3T MRI devices cost over a million dollars and can service approximately 2000-3000 patients a year. Under this model, it is not economically feasible to provide service to the one billion patients with neurological disorders worldwide. In Chapter 4, I review recent advances in software and hardware that have enabled the development of lower-cost MRI devices that provide reasonable image quality while using cost-effective low-field
magnets. Since the start of my PhD, at least five new low-field devices have received FDA clearance, including the device I studied in Chapter 5 and 6. Determining the appropriate clinical use cases and delineating what these novel devices are and are not capable of will be an essential next step. Additionally, there remain significant challenges in deploying technology in areas that may not have personnel with technical training in device operation, maintenance, and interpretation. Machine learning may provide some assistance in smoothing the transition as medical imaging access is scaled up. However, algorithms cannot obviate the need for thoughtful device design and investments in technical training for local stakeholders.

The primary objective of the work in this thesis was to increase access to medical imaging options for patients. To approach this problem, in Chapter 2 and 3 I developed automated methods for analyzing postoperative epilepsy images. The postoperative epilepsy patient population is understudied, as working with postoperative images is difficult and patients are often lost in followup. I hope these automated tools for analyzing postoperative data will spur research in this population, thus providing clinicians with an understanding of how surgery impacts the brain and promoting better care options for epilepsy patients. Chapters 4, 5, and 6 focused on clinical applications of portable, low-field MRI devices. These devices have the potential to dramatically lower imaging costs and make MRI devices accessible in remote and lower-resource settings. Encouragingly, device manufactures are already using the code from Chapter 5 to simulate data for their own low-field devices. Additionally, the work in Chapter 6 has motivated the formation of a low-field MRI working group within the North American Imaging in MS Cooperative. I hope that the work outlined in this thesis will continue to contribute to the development and deployment of lower-cost MRI devices across the world. To aid in clinical translation of the work, open access to all code related to these research projects has been provided on GitHub.
REFERENCES


54. Morgan VL, Rogers BP, Anderson AW, Landman BA, Englot DJ. Divergent network properties that predict early surgical failure versus late recurrence in temporal lobe


75. Fernandes DA, Yasuda CL, Lopes TM, et al. Long-term postoperative atrophy of contralateral hippocampus and cognitive function in unilateral refractory MTLE with


111. Moser E, Laistler E, Schmitt F, Kontaxis G. Ultra-High Field NMR and MRI—The Role


Source Hardware – Case Study of an Open Source Magnetic Resonance Imaging Scanner.


158. Mills TT. FDA Clearance K192002 - Lucy Point-of-Care Magnetic Resonance Imaging Device. FDA. 2020;


192. Jimeno MM, Vaughan JT, Geethanath S. Superconducting magnet designs and MRI accessibility: a review. 2022;


196. Diehl JC, Van Doesum F, Bakker M, et al. The embodiment of low-field MRI for the


224. Kim EH, Oh MC, Kim SH. Application of low-field intraoperative magnetic resonance imaging in transsphenoidal surgery for pituitary adenomas: Technical points to improve


230. Hall WA, Kowalik K, Liu H, Truwit CL, Kucharczyk J. Costs and benefits of


277. Watchmaker JM, Xia D, Dayan E, et al. Point-of-Care Knee Imaging Using a 64mT Portable MRI Scanner: First Results. ISMRM Low-Field MRI Workshop; 2022.


279. Guerini H, Omoumi P, Guichoux F, et al. Fat Suppression with Dixon Techniques in


299. Chollet F. Keras. 2015.


308. Yoo Y, Tang LYW, Brosch T, et al. Deep learning of joint myelin and T1w MRI features in normal-appearing brain tissue to distinguish between multiple sclerosis patients and


337. Boudraa AO, Dehak SMR, Zhu YM, Pachai C, Bao YG, Grimaud J. Automated


180

