

Advanced Methods for Robust Neuroimage Analysis

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Clemens Pollak
aus
Hannover, Deutschland

Bonn, 15.12.2025

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der
Rheinischen Friedrich-Wilhelms-Universität Bonn

Gutachter/Betreuer: Prof. Dr. Martin Reuter
Gutachter: Prof. Dr. Reinhard Klein

Tag der Promotion: 13.07.2026
Erscheinungsjahr: 2026

Abstract

Magnetic resonance imaging (MRI) is the central modality for acquiring images of the brain in biomedical research. Advanced computational pipelines aid researchers along the way, from initially acquiring MRI to ultimately reaching research conclusions. The researchers’ interaction with these pipelines impacts the power and validity of research findings. Software tools can mitigate error and bias, but inadequate analysis methods can also introduce them. To aid the research community, we propose three novel methods that make neuroimage analysis more robust and intervene at different points in the analysis process.

First, we propose a method to track head motion during the MRI acquisition. Motion artifacts are the most prominent MRI artifacts and are known to cause biased morphometric analysis results. Our measurements allow researchers to reduce this bias by including head motion in statistical modeling. The method introduces an improved robust registration approach for depth images acquired by an optical tracking camera, along with head motion metrics and novel evaluation approaches. Additionally, we show that the acquired measurements can be used to train a deep-learning model that estimates the head motion from images alone. Second, we address a major challenge in the robustness of neuroimaging methods by providing a novel approach for analyzing lesion-affected scans. We propose a lesion inpainting tool that replaces damaged tissue with healthy-looking tissue in MRI scans. This tool allows consecutive analyses to run as if no lesions were present, enabling new analyses on data that would previously be discarded. It leverages a denoising diffusion probabilistic model (DDPM) combined with a novel inference scheme, allowing for mask-independent, memory-efficient volumetric inpainting. Finally, we address multiple gaps in current corpus callosum analysis methods. The corpus callosum (CC) is the central connection between the brain’s hemispheres and of high interest for research of disease and aging. Previous work has successfully developed automated tools for CC analysis but has shortcomings in accuracy, reliability, and performance. To close this gap, we propose a comprehensive corpus callosum analysis tool that performs segmentation and surface reconstruction and then derives suitable metrics for statistical modeling. We show that building a

state-of-the-art analysis method extends far beyond simple volumetric segmentation. Our method standardizes the head position specifically for unbiased segmentation, then uses the CC segmentation to derive curvature, length, thickness, and other advanced CC-specific metrics. The combination of method improvements in every step results in a method that detects disease effects where they would be missed by previous methods. We place special attention on the thorough evaluation of the proposed methods, testing generalization to unseen datasets and diseases. All proposed methods were released to the research community as open-source tools (on github.com/Deep-MI). The motion-tracking tools are currently deployed in the Rhineland Study, a large population study with over 10,000 participants. Furthermore, the general image processing tools are integrated into the established FastSurfer neuroimage toolbox, allowing the global research community to seamlessly interface with these novel analysis techniques.

Acknowledgments

I would like to express my foremost gratitude to Prof. Dr. Martin Reuter for his continuous guidance and support during my years in his group. Furthermore, I am grateful to David Kügler for his unrelenting support and trust. I would like to thank Kersten Diers for being a calm and always positive influence in my hectic research journey. I also sincerely thank Christian Ewert for inspiring discussions, companionship, and dragging me out of the tunnel of debugging many times. I am also grateful to have Santiago Estrada as a colleague, showing me new points of view on matters in research and beyond. Special thanks also to Jingjing Wu and Sebastian Rassmann, who brought new perspectives to my research endeavors.

I also appreciate the support and advice from my collaborators and their research groups, namely Prof. Dr. Theodor Rüber, who brought new clinical perspectives into my work, and Prof. Dr. Monique M.B. Breteler, whose feedback and vital support in data acquisition helped put my PhD on track. I am grateful to Prof. Dr. Markus Schirmer, who showed me a new research environment at the Massachusetts General Hospital, where I truly expanded my horizon. Special thanks also to Prof. Dr. Reinhard Klein from the University of Bonn for agreeing to supervise my thesis and dedicating time to review my work.

Finally, I want to thank all my friends in and outside of the workplace who made this journey truly special. Most importantly, I would like to thank my parents, Ulrike and Karl-Heinz, for their unwavering support and belief in me and my wife, Shu Farn, for paving the way for the start, end, and all the steps in between.

Contents

Abstract	4
Acknowledgments	6
1. Introduction	1
1.1. Motivation	1
1.2. Objectives	2
1.3. Contributions	2
1.4. Organization	4
2. Background	5
2.1. Brain morphometry studies	5
2.1.1. MRI acquisition & reconstruction	6
2.1.2. Image standardization	9
2.1.3. Feature extraction	13
2.1.4. Statistical testing	14
2.1.5. Contributions in the context of neuroimaging studies	16
2.2. Technical terms and definitions	17
2.2.1. Deep learning methods	17
2.2.2. Robust estimators in registration	18
2.2.3. Evaluation metrics	19
2.2.4. Statistical testing	21
3. Quantifying head motion	22
3.1. Summary	23
3.1.1. Motivation	23
3.1.2. Contributions	23
3.1.3. Motion tracking pipeline	24
3.1.4. Evaluation	25
3.1.5. Results	26
3.1.6. Conclusion	26
3.2. Author Contributions	27

4. Estimating head motion from MR-images	28
4.1. Summary	29
4.1.1. Motivation	29
4.1.2. Contributions	29
4.1.3. Motion estimation pipeline	29
4.1.4. Evaluation	30
4.1.5. Results	31
4.1.6. Conclusion	31
4.2. Author Contributions	31
5. Lesion inpainting tool for whole-brain MRI segmentation	32
5.1. Summary	33
5.1.1. Motivation	33
5.1.2. Contributions	33
5.1.3. Lesion-robust whole brain analysis pipeline	34
5.1.4. Evaluation	35
5.1.5. Results	35
5.1.6. Conclusion	36
5.2. Author Contributions	37
6. A robust, accurate and comprehensive framework for corpus callosum morphometry	38
6.0.1. Motivation	39
6.0.2. Contributions	40
6.0.3. Overview of the corpus callosum morphometry pipeline	41
6.0.4. Evaluation	42
6.0.5. Results	42
6.0.6. Conclusion	43
6.1. Author Contributions	43
7. Summary and Outlook	44
7.1. Head motion	45
7.2. Whole-brain analysis in the presence of lesions	45
7.3. Robust corpus callosum analysis	46
7.4. Future work	46
7.5. Conclusion	48
A. Publication: "Quantifying MR head motion in the Rhineland Study – A robust method for population cohorts"	50
B. Publication: "Estimating Head Motion from Mr-Images"	67

C. Publication: "FastSurfer-LIT: Lesion inpainting tool for whole-brain MRI segmentation with tumors, cavities, and abnormalities"	75
D. Publication: "FastSurfer-CC: A robust, accurate and comprehensive framework for corpus callosum morphometry"	101
Bibliography	126
Bibliography for the publication: Quantifying MR head motion in the Rhineland Study – A robust method for population cohorts	137
Bibliography for the publication: Estimating head motion from MR-images .	147
Bibliography for the publication: FastSurfer-LIT: Lesion Inpainting Tool for Whole Brain MRI Segmentation with Tumors, Cavities and Abnormalities	150
Bibliography for the publication: FastSurfer-CC: A robust, accurate, and comprehensive framework for corpus callosum morphometry	159
List of Publications	168
List of Figures	169

Chapter 1.

Introduction

1.1. Motivation

Robustness in a software context can be defined as software being "capable of performing without failure under a wide range of conditions" [1]. While robust software methods are desirable in general, software used in a medical context requires utmost care in its design, testing, and application to minimize negative effects on patient care and biomedical research. Specifically, neuroimaging research is experiencing a rapid growth in available methods, some of which are distributed to the global research community as readily available analysis tools. When selecting an analysis method, robustness is a critical but multifaceted criterion. Foremost, robust methods are trustworthy and can produce valid results without extensive modification or continuous validation. Furthermore, more reliable measurements can reveal previously missed correlations from the same body of measurements. They can also allow researchers to reach the same conclusions using fewer measurements, which can ultimately lead to fewer required participants in clinical trials and research studies, reducing research costs and the burden on medical professionals. Finally, robust methods may be applied to previously underrepresented groups, like individuals with rare diseases, reducing selection bias and ultimately improving fairness in medical imaging.

Unfortunately, the diversity of cohorts, MRI scanners, and scan protocols makes it challenging to evaluate method robustness. Even the most common sources of error and bias, like image quality and anatomical variation, are challenging to fully capture in large multi-site datasets. While the high heterogeneity of neuroimaging data is already a grand challenge, some sources of bias and confounders are specific to individual brain structures or analysis methods. Therefore, no general solution to medical imaging robustness can exist, but robustness has to be built into study design, method development, and analysis individually. In this thesis, we propose three different approaches to tackle three major challenges in neuroimage robustness, which, combined with rigorous testing, showcase a general set of paradigms to address robustness concerns through method innovation.

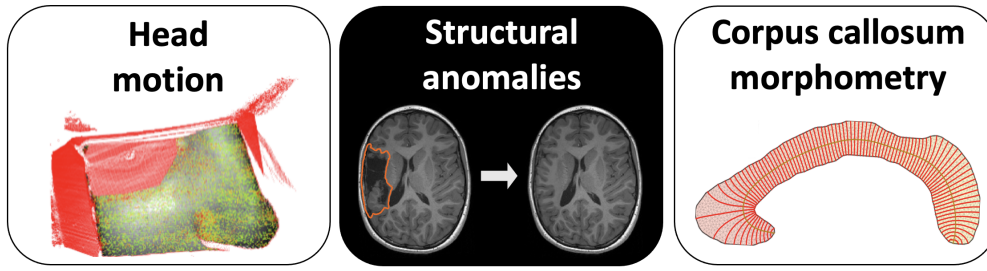


Figure 1.1.: Overview of the three challenges addressed in this thesis.

1.2. Objectives

The overarching objective of this work is enabling accurate and reliable brain morphometry based on structural MRI. Three major challenges in common neuroimage pipelines are:

- Head motion artifacts causing analysis errors and biased morphometry
- Brain lesions biasing or completely preventing automated morphometric analysis
- Hidden analysis errors due to erroneous standardization of the head pose, segmentation errors, and choice of analysis metrics, for example, in the analysis of the corpus callosum.

This thesis addresses these challenges with methodological improvements while carefully selecting appropriate points of intervention in the established neuroimage analysis workflow.

1.3. Contributions

We propose a suite of methods to address each of the aforementioned challenges:

- **Measuring head motion:** First, we propose a set of methods to estimate head motion during the MRI acquisition based on recordings from an in-MRI depth camera. After extensively validating the method, we use it to train a deep learning model that predicts average head motion directly from an MR image. These estimates can be included as a confounder in statistical modeling, provide automated quality control, and enable further research on strategies to mitigate head motion.

- **Inpainting brain lesions:** To enable processing of brain MRI with large lesions, e.g., post-surgery or radiotherapy, we develop a general lesion inpainting tool that replaces damaged tissue with healthy-looking tissue in the MR image. Then we combine this novel tool with a modified whole-brain segmentation and cortical surface reconstruction pipeline. Our method results in segmentation accuracy comparable to lesion-free MRI, despite none of the involved methods being trained on images with lesions.
- **Corpus callosum morphometry:** Accurate and unbiased corpus callosum analysis in structural MRI relies on multiple interdependent factors. First, accurate standardization of the head pose is critical for choosing an unbiased frame of reference for later analysis. Second, robust and well-validated segmentation tools are required for reliable volumetric descriptions. Third, advanced shape metrics are needed to measure disease-specific changes. Since one or multiple of these criteria are often overlooked in prior work, we develop a fast, accurate, and comprehensive corpus callosum morphometry framework, which measures, for example, corpus callosum sub-regions, curvature, length, and local thickness. Our tool's accuracy remains stable in challenging scenarios, finds previously missed correlations, and addresses shortcomings of prior corpus callosum sub-segmentation schemes.

In addition to methodological contributions, a notable aspect of our work is the comprehensive validation performed to ensure that tools are ready to use by the research community. Where appropriate, we employ datasets unseen during development, realistic use cases, and deliberate out-of-distribution testing to rigorously evaluate a method's performance. Finally, we release the tools as software packages (github.com/Deep-MI/), making them directly available to the research community. The methods, along with their evaluations, are extensively discussed in four publications:

- **Clemens Pollak**, David Kügler, Monique M. B. Breteler & Martin Reuter (2023). "Quantifying MR head motion in the Rhineland Study – A robust method for population cohorts". *NeuroImage*, 275, 120176.
DOI: [10.1016/j.neuroimage.2023.120176](https://doi.org/10.1016/j.neuroimage.2023.120176)
- **Clemens Pollak**, David Kügler, & Martin Reuter (2023). "Estimating head motion from MR-images". 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), pp. 1-5. IEEE
DOI: [10.1109/ISBI53787.2023.10230717](https://doi.org/10.1109/ISBI53787.2023.10230717)
- **Clemens Pollak**, David Kügler, Tobias Bauer, Theodor Rüber, & Martin Reuter (2025). "FastSurfer-LIT: Lesion inpainting tool for whole-brain MRI segmentation with tumors, cavities, and abnormalities.". *Imaging Neuroscience*, Volume 3 (2025).
DOI: [10.1162/imag_a_00446](https://doi.org/10.1162/imag_a_00446)

- **Clemens Pollak**, Kersten Diers, Santiago Estrada, David Kügler, & Martin Reuter. *"FastSurfer-CC: A robust, accurate, and comprehensive framework for corpus callosum morphometry"*. arXiv, currently under review at Imaging Neuroscience.
DOI: [arXiv.2511.16471](https://arxiv.org/abs/2511.16471)

1.4. Organization

This thesis is divided into the following five sections:

- Chapter 2 presents general background information on neuroimaging studies.
- Chapter 3 presents a summary of our publication "Quantifying MR head motion in the Rhineland Study – A robust method for population cohorts" [2].
- Chapter 4 presents a summary of our publication "Estimating head motion from MR-images" [3].
- Chapter 5 presents a summary of our publication "*FastSurfer-LIT*: Lesion inpainting tool for whole-brain MRI segmentation with tumors, cavities, and abnormalities" [4].
- Chapter 6 presents a summary of our publication "*FastSurfer-CC*: A robust, accurate, and comprehensive framework for corpus callosum morphometry" [5].
- Chapter 7 concludes the thesis by discussing the impact of the presented methods and providing an outlook of potential future directions.

Chapter 2.

Background

This chapter first outlines the workflow of a typical brain morphometry study, commencing with image acquisition and concluding with statistical analysis to answer the research question. This introduces key points of vulnerability (sources of error and bias) that motivate the methodological contributions in the following chapters. It will also showcase the heterogeneous landscape of neuroimaging data, which outlines the range of conditions the developed tools need to operate in. The second section of this chapter presents technical background information about methods, evaluation criteria, and statistical testing introduced in the later chapters.

2.1. Brain morphometry studies

For simplicity and brevity, we will focus on brain morphometry studies using structural MRI, e.g., researching whether a diseased population shows structural differences from a general population. In current studies the setups are frequently more complex, involving, for example, multiple MRI modalities, genomics, and clinical data about patient outcomes. Despite this complexity, structural analysis of the brain is the basic building block for many other more complex analyses. Most steps in this structural analysis are standardized using established software toolboxes and procedures, which are utilized in studies across the globe. In Figure 2.1 we divide the process into four steps. Since this thesis is focused on image analysis, we omit the study planning and participant recruitment stages and start with (i) MRI acquisition, where the brain is measured by an MRI scanner, resulting in a 3D image (MRI volume). The next step (ii) is a preprocessing step, where images in a study population are standardized and prepared for (iii) feature extraction. The feature extraction step derives morphometrics from the standardized images. Often neuroimaging toolboxes provide a plethora of metrics and intermediate outputs, like segmentations and structure sizes, surface

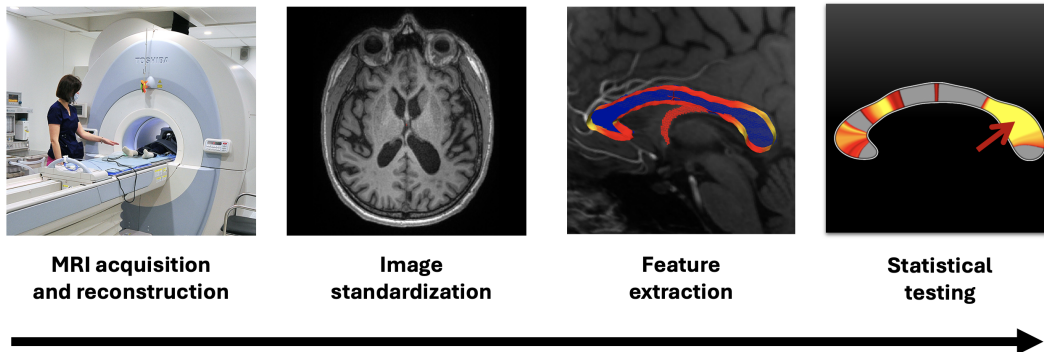


Figure 2.1.: Overview of the discussed stages of a structural MRI study.

The MRI photo is courtesy of Wikimedia commons, author Mos.ru, Vantage Atlas MRI scanner from Toshiba, author Mos.ru, CC-BY 4.0

models, and surface thickness. Then in a final step (iv), the researcher confirms or denies the research hypothesis by selecting appropriate metrics and performing appropriate statistical tests. In the next sections, we will discuss these steps in more detail and uncover sources of error, which can propagate to the subsequent stages.

2.1.1. MRI acquisition & reconstruction

This section describes the process of acquiring an MRI scan, from selecting appropriate imaging protocols to finally saving the MR image for later analysis.

Scan preparation

The MRI acquisition is sometimes understood to be the protocol provided to the MRI scanner to perform its measurements. A complete view of the image acquisition, however, starts with the study participant. In most studies participants provide informed consent, are checked for contraindications, and can then enter the room with the MRI scanner. Even in these initial stages, interventions can help with MR-image quality. Children, for example, can receive a brief mock-scan training [6], and claustrophobic participants can be helped to prepare for the scan with relaxation techniques and familiarizing themselves with the environment [7]. These interventions can reduce the scan abortion rate and in-MRI motion of participants. After participants lie down on a movable scan table and place their head on a recessed pillow, padding can be applied around the head, and headphones as well as a screen for movies can be set up

to further increase patient comfort and reduce head motion and the resulting motion artifacts [8–10]. To measure the head motion during the scan, an external camera [11] can be positioned on the head coil, which we will discuss extensively in the next chapter. After the participant is positioned comfortably, the head coils are placed over the participant’s head, and the MRI technologist will move the table and mark an anatomical landmark (e.g., chin or sternum) with a laser. Then the patient’s head is automatically moved to the center of the main magnet, and the MRI scanner can begin imaging the participant’s brain.

MRI scanner

MRI is the gold standard for structural imaging of the brain. Therefore the MRI scanner is the central tool for acquiring neuroimaging study data. The largest part of an MRI scanner is the main magnet. Depending on the scanner, this magnet can have field strengths ranging from less than 0.01 Tesla (ultra low-field MRI [12]) to more than 11 Tesla (ultra high-field MRI [13]). Most common are scanners with field strengths of 1.5 or 3 Tesla, with higher field strengths generally offering higher fidelity images, and ultra-low field MRI offering smaller and even portable scan setups – different from most current studies. The scanner itself is a large source of heterogeneity in multi-site neuroimaging studies. Scanners from different manufacturers can produce images with different fidelity and contrast, even when the acquisition parameters are standardized across scanners [14]. Additionally, image quality can vary over time even using the same MRI scanner [15], requiring constant quality control for consistent imaging.

MRI modalities

During MR imaging the head coils are turned on and off to excite molecules and then measure their relaxation time. Depending on the selected MRI protocol, different molecules are excited, or measurements are taken in different orders or densities. The MRI protocols are complex instructions for the MRI scanner and can drastically and subtly vary the MR image. Exciting different molecules gives images a different contrast, which is considered a distinct MRI modality. For example, fat can appear bright and water can appear dark (T1-weighted sequence), or both can appear bright (T2-weighted sequence), or scanners can add a timing pulse to null the signal from normal free fluid while keeping T2w-contrast (FLuid Attenuated Inversion Recovery, FLAIR). Other parameters affect the quality and resolution of the resulting image, with longer sequences usually providing better image quality. Robust MRI sequences are an active field of research, with many subfields. Even the order of measurements, for example, affects the robustness of the sequence to motion artifacts [16]. After appropriate MRI

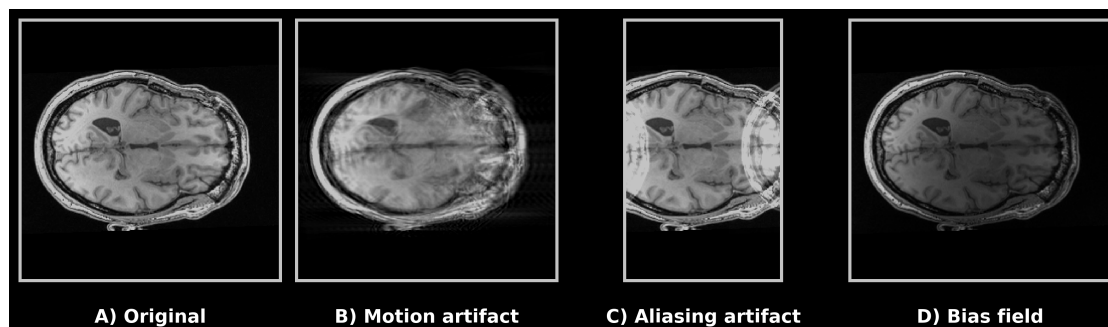


Figure 2.2.: Simulations of strong MRI artifacts for three different artifact types (B-C) with the original image (A) for comparison.

sequences are selected, the scanner performs calibration and automatically sets the field of view while supervised by the MRI technologist. Finally, the MRI scanner measures the participant’s brain by recording the molecule relaxation pulses according to the selected MRI modality.

MRI reconstruction

The scanner’s measurements result in a (sparse) image in the frequency domain. To turn the frequency measurements into a (magnitude) grayscale image, the scanner executes a proprietary reconstruction algorithm. These algorithms can already include steps that we may originally attribute to the next section of image standardization, e.g., reduction of motion artifacts, bias field correction, and up-sampling based on sparse measurements [17]. After the reconstruction is completed, the original measurements are discarded along with the imaginary part of the reconstruction. The magnitude MR image is saved along with metadata and can be processed in hospital PACS (picture archiving and communication) systems or study databases. Researchers can then access the images along with metadata to perform image standardization and analysis.

Pitfalls and limitations

Errors during the acquisition most prominently present themselves as MRI artifacts. The most prevalent type of artifacts are motion artifacts [18] caused by head motion during the acquisition (see Figure 2.2B). Motion artifacts can consistently alter, e.g., estimates of cortical thickness in later analysis [19]. While increasing patient comfort can significantly reduce head motion, other subconscious movements like respiration

cannot be completely suppressed. Besides head motion, common artifacts are aliasing artifacts, which occur when the field of view is inaccurate or too small for the head. Here, out-of-view parts of the head wrap around to the other side of the image (e.g., the nose appearing at the back of the head). This is challenging to correct for retrospectively but can be addressed with various MRI sequence modifications, like oversampling, increasing the field of view, and changing the phase encoding direction. The bias field is another type of artifact, showing as intensity non-uniformity in the MRI, based on imperfections in the magnetic field. The bias field is mainly caused by the interaction of the magnetic field with the head but can also be a sign of poor scanner calibration or magnet performance. Retrospective correction is, however, a standard tool in many neuroimaging toolboxes [20–22]. Another major challenge in neuroimaging is the data heterogeneity. In multicenter studies, participants may be scanned using various scanners and protocols. Ideally, analysis tools should function independent of the MRI scanner used to image the participant. Finally, practical considerations, like the participant’s body size, may amplify errors or cause selection bias. Obese people, for example, can sometimes not lie in the MRI comfortably or can not enter the scanner at all – leading to their exclusion. Similar issues can arise for people with claustrophobia or motion disorders. Ultimately researchers cannot address all errors and biases in brain MRI studies, but by designing subsequent methods to handle errors and heterogeneity, we can mitigate their effects.

2.1.2. Image standardization

To reduce the aforementioned data heterogeneity, most feature extraction requires prior image standardization. Common standardization steps can be divided into six categories:

- Intensity standardization
- Standardization of head pose
- Standardization of image resolution
- Artifact removal
- Exclusion
- Advanced standardization and harmonization

This list is not exhaustive and specifically does not include highly specific standardization steps, like cropping the image to show a specific structure, since those steps exploit knowledge about the brain structure and therefore have overlap with the next stage of feature extraction. Since the general steps in these lists can be skipped and reordered arbitrarily, depending on the standardization methods of choice, we discuss each step separately in the following sections.

Intensity normalization

Intensity normalization is the most straightforward step of the standardization stage. Since most MRI modalities do not provide absolute measurements but define the contrast between brain regions, the intensity profiles across images can differ drastically. The goal of intensity normalization is to unify intensity profiles while keeping the participant-specific contrast differences. For deep-learning methods intensities are usually expected to be between zero and one or centered around zero with a standard deviation of one. Traditional tools [23] sometimes scale to 8-bit integer representation to save computational resources. While a simple scaling is sometimes sufficient, some voxels may appear extremely bright due to measurement errors. Therefore, a robust rescaling can provide more uniform intensity profiles. A common approach is to set a percentile of the highest values to the maximum value of the scaling range and then scale the other values as before. Other more sophisticated approaches, like scaling the gray or white matter to predefined intensity ranges, can even introduce meaningful absolute intensities into standard MRI.

Standardization of head pose

While the MRI technologist likely already adjusted the scan field of view of the scanner to fit the participant's head position, this adjustment is often not accurate or reliable enough for consecutive analyses. Therefore, the head position is often standardized with registration to a template or, as we propose in Chapter 6, with anatomical landmarks. Registration-based approaches consider global information for the alignment, making them potentially robust to local changes. Landmark-based standardization, on the other hand, can produce unexpected results when landmarks are distorted/missing in abnormal anatomies, but they offer verifiable, anatomically motivated alignments, which are crucial for some applications. Often they are also faster than traditional registration approaches. Changing the head pose in an image standardizes it but also degrades the image quality due to interpolation artifacts. Therefore, methods that are robust to changes in head position can avoid the standardization of head pose and therefore leverage more accurate information. They can also generate intermediate volumetric outputs (e.g., segmentations) in the desired output space, avoiding interpolation after the feature extraction as well.

Standardization of image resolution

Image interpolation is also frequently used to standardize the image resolution – also known as down-sampling and up-sampling. In addition to interpolation artifacts, image

information is discarded when images are of higher resolution than the desired standard (down-sampling). Methods that can operate on the original resolution can therefore use more information and produce higher-quality analysis results – making methods that operate natively on multiple resolutions highly desirable. For deep-learning methods this was introduced by FastSurferVINN [24] using interpolation layers that standardize the image resolution only in the latent space of the network. This allows training and inference on multiple resolutions natively, allowing the networks to learn from high- and low-resolution data at the same time.

Artifact removal

MRI artifacts are highly specific to the MRI scanner and the participant. Motion artifacts, for example, are almost universally present in some populations (e.g., children [25]), while other, highly compliant, populations exhibit minimal head motion, which is dominated by respiratory movement and barely visible [3]. Additionally, the manifestation of motion artifacts is highly dependent on the MRI scanner and protocol [26], with some sequences specifically being designed to be robust to head motion or to mitigate its effects [16]. Therefore, retrospective removal of motion artifacts is challenging, especially without information about the MRI reconstruction. A more consistent artifact is the scanner’s bias field. The bias field still depends on both the participant and the scanner, but it typically exhibits a similar pattern, with scanners of higher field strength generally displaying a stronger bias field. A popular tool for this task is N4 bias field correction [27], which iteratively deconvolves the intensity histogram with a Gaussian, remaps the intensities, and spatially smooths the output using a B-spline model of the bias field. This is based on the assumption that the low-frequency bias field corruption can be represented by a Gaussian convolution of the intensity histograms. There are many more types of artifacts (see e.g., [28] for an overview), with dedicated mitigation and correction strategies, though most of them are typically addressed during image acquisition and reconstruction and don’t have to be tackled by researchers retrospectively.

Exclusion

Most MRI studies have multiple stages of quality control, where images are flagged or excluded. The first check happens directly after the scan, when the MRI technologist views the image to ensure the scan was completed successfully. At this stage, the participant can be scanned again – e.g., in case strong motion artifacts are present or acquisition parameters were incorrect. Prior to image analysis, the MRIs are usually

reviewed by researchers, who ensure sufficient quality to answer their research question. This process can be quite rigorous and can target specific areas, excluding, for example, participants with brain injury or local artifacts. It can also be mandated to screen for incidental findings. Finally, intermediate analysis results, like segmentations and surfaces, can be reviewed along with analysis metrics to ensure that the feature extraction has succeeded. While all quality control steps are likely to reduce errors, the second step of rigorous review is the most impactful. Here decisions about inclusion and exclusion can be made with ample time and with the research question in mind. Only reviewing analysis outputs may seem compelling, but many methods will crash or not warn the user and instead produce realistic-looking outputs that have no basis in valid measurements. Beyond simply excluding problematic cases, researchers can attempt to recover some data (e.g., by artifact removal or correcting metadata), but this individual treatment is challenging in large studies.

Advanced standardization and harmonization

Beyond the previously discussed standardization methods, there are more complex approaches that can further unify datasets for neuroimage analyses. One prominent example is the image harmonization tool SynthSR [29], which performs contrast adjustment that can turn other MRI modalities or CT images into synthetic structural MR images of one millimeter resolution. Harmonization tools generally aim to remove non-biological variance in MRI, often combining previously discussed standardization with additional (often learned) adjustments to the appearance of the brain in the MRI. In Chapter 5 we will introduce an advanced standardization method that focuses on lesion in-painting. This process replaces damaged areas with healthy-looking areas in the MRI to enable running tools that usually cannot handle damaged tissue (lesions). Some advanced standardization and harmonization techniques risk altering biological variance in the image, which can potentially affect the validity of research findings. Therefore, thorough evaluation is required.

Pitfalls and limitations

Data standardization is an essential step in current neuroimage analysis pipelines. Sometimes the choice of standardization is done explicitly by researchers performing the analysis; other times they are chosen by method developers and directly integrated with feature extraction. Nonetheless, understanding how the analysis process alters the image information is crucial. In the prior subsections, we saw that many standardization approaches are accompanied by information loss, which can reduce accuracy or introduce errors and bias in subsequent steps. Additionally, some standardization steps,

like traditional registration and bias field removal, can add considerable processing time. Therefore, less standardization is preferable, but only if the following methods are robust enough to still produce high-quality features for statistical analysis.

2.1.3. Feature extraction

During the feature extraction stage, measurements are taken from the MR image. The tools at this stage are often highly specific to a structure or even a research question. Nonetheless, many current structural image feature extraction tools share a set of similar output formats and paradigms.

Intermediate outputs

Most methods first extract an intermediate output that describes the structure in question. In whole brain analysis toolboxes the most prominent intermediates are segmentations of individual structures and surface models, e.g., of the cortical surface [21, 30]. These intermediates can be quality controlled and evaluated with respect to accuracy, reliability, and robustness. Then metrics like the structure's volume or the surface thickness are calculated. Intermediates improve the interpretability and evaluation of methods but are not available for all markers. Brain age difference [31], for example, is a popular biomarker that describes the difference between the predicted age of a participant based on brain MRI and the chronological age. This more abstract marker can be predicted directly from the image, in which case no intermediates are present for verification or inspection.

Morphometrics

Researchers seek highly descriptive metrics for statistical evaluation to describe the brain structure (morphology). A selection of these morphometrics should be able to answer the research question by summarizing complex structural changes in a small set of values. A common and straightforward metric is the volume of a structure, which can often suffice to explain group differences. However, complex brain changes can be missed by simple volumetric descriptors. Changes in different regions of a structure can, for example, be associated with different diseases [32]. Additionally, structural changes in the brain can cause complex pressure changes, causing changes in a structure's shape without visible loss of tissue. Therefore, advanced but interpretable shape descriptors, like thickness or curvature, can be required to capture structural differences in neuroimaging studies. Ultimately, the choice of metrics is dependent on the research question, which is why neuroimaging tools and toolboxes offer a wide range of metrics for researchers to choose from.

Pitfalls and limitations

The variability among human medical images means that currently no methods can be applied to any MRI without adjustments. Due to genetic variance or surgery, entire structures of the brain can be missing or change their positions considerably. Many of these extreme changes are rare but can be of high interest for research. Frequently, researchers need to perform rigorous evaluation to gauge whether a method is applicable to their research question and which level of quality control is required. When a new method is developed, this evaluation should already be performed for a set of common (and ideally also some uncommon) scenarios. When intermediate outputs are generated, for example, these can be compared to manual annotations or other methods' outputs. The final metrics can be used to reproduce known effects, testing whether increased accuracy and reliability lead to lower p-values. Performing these tests in a realistic environment, using unseen data, allows the research community to judge whether the method will perform adequately on a dataset found in the wild. Ideally the method's generalizability is also tested during evaluation by including, e.g., disease or MRI protocols that were not part of the original dataset. In cases where training and evaluation are limited to a single homogeneous dataset, evaluations could be skewed by valuing accuracy over generalizability – essentially encouraging overfitting. A detrimental error in the evaluation of methods is *data leakage*, which describes generating overly optimistic results by using data in the training process that is also used during evaluation. Generally errors by the feature extraction method are more challenging to predict than in previous stages, especially since many tools rely on deep learning, where interpretability is still an open research question. Nonetheless, best practices like sanity checking and defensive programming can often raise errors when they occur and ease the burden of manual quality control.

2.1.4. Statistical testing

While advanced statistical modeling of biomedical data is not a focus of this thesis, it is a crucial step in neuroimaging analysis. Here researchers combine appropriate features with appropriate statistical testing to discern the significance of group differences or longitudinal effects.

Hypothesis-driven analysis

In hypothesis-driven analysis, researchers aim to confirm or deny their research hypothesis, using features specific to the research question – often changes in a specific

structure or brain region. Often multiple metrics are available that describe slightly different structural characteristics but are highly correlated. Furthermore, researchers have to identify confounders of their analysis (like the intracranial volume or age) and also include them for appropriate statistical modeling. Multiple features can be included, e.g., using general linear models (GLM) [33], defined as

$$Y = X\beta + \epsilon, \quad (2.1)$$

where Y is the structural measure, X are the predictors, β represents the effect size, and ϵ is the residual. In this modeling, however, correlation between measures may reduce their individual significance. Therefore, it is advisable to carefully select only the most robust descriptors, thereby maximizing statistical power and interpretability while mitigating issues of multicollinearity.

Exploratory and mass-univariate analysis

In contrast to hypothesis-driven approaches, exploratory, data-driven analyses do not restrict the search to a predefined region. Instead, they assess the entire brain to uncover spatially distributed patterns of atrophy or development. Often mass-univariate frameworks are used for this purpose. In this approach, general linear models are applied independently to a large body of measurements, sometimes across all image voxels or vertices of a cortical surface. While this can allow researchers to discover new, unexpected correlations, it also drastically increases the chance of false positives. Therefore, multiple comparison correction methods need to be applied before interpreting the results.

Pitfalls and limitations

Any type of advanced statistical analysis comes with its pitfalls. One prevalent issue is "double dipping" [34]. This describes a circular hypothesis-driven analysis, where data used to select a statistically significant region is subsequently used for the final statistical testing. Since the noise in the selected region aligns with the selection criteria, analysis results will be distorted and potentially invalid. Another common set of errors are assumption violations. Longitudinal modeling, for instance, is another common type of statistical modeling, where brain changes are observed over time. Here the GLM can not be used as described above, since the residual errors are not independent. An appropriate statistical model is, for example, linear mixed effect models (LME) [35],

$$Y = X\beta + Zu + \epsilon, \quad (2.2)$$

with the term Zu describing the random effects design matrix and corresponding random intercept and slope. Overall, appropriate statistical modeling and potential errors are specific to the research question, but following best practices in study design, model selection, and validation can avoid most pitfalls.

2.1.5. Contributions in the context of neuroimaging studies

The methods and resulting tools presented in this thesis improve the robustness of structural neuroimage analysis at carefully chosen intervention points in the presented study process. The first presented projects around head motion are interlinked with the scan process. Here, we measure head motion directly at the scanner to provide a confounding variable for the final statistical analysis step. But we also formulate the motion estimation task as a feature extraction task, where we predict head motion during the acquisition from MRI images alone. The second project makes analysis robust to brain lesions with a novel standardization method and then combines it with suitable pre-existing feature extraction methods. Making the feature extraction itself robust may be appealing at first, but the lack of ground truth labels for deep learning and the requirement to modify every feature extraction tool in a large toolbox make this approach much less effective. Finally, we combine head pose standardization and novel feature extraction to derive robust metrics for statistical analysis in our corpus callosum analysis framework.

2.2. Technical terms and definitions

This section briefly introduces some core concepts related to the later discussed deep learning and registration methods, evaluation criteria, and statistical testing. More detail is provided in the method sections of the corresponding publications.

2.2.1. Deep learning methods

The **Convolutional Neural Network** (CNN) is a basic neural network architecture, which uses filter kernels (convolutions) for (image) processing.

The **DenseNet** [36] builds on the early CNN architecture, which introduces blocks of densely connected convolutions. Here, one convolution receives the outputs of all previous convolutions in a block as input. This, among other benefits, addresses the vanishing gradient problem and allows for training of deeper networks.

Fully Convolutional Neural Network (FCN) [37] architectures use only convolutions and specifically omit, e.g., fully connected layers often used in standard CNNs. This architecture uses few parameters for deep models and is expected to focus on local features.

The **U-Net** [38] is a type of neural network specifically designed for image segmentation and is based on FCNs. Inside the U-Net, input data is first down-sampled and later up-sampled again, typically producing the same size output as was given as input.

The **Voxel size Independent Neural Network** (VINN) [24] builds on the U-Net architecture and can natively handle inputs at various image resolutions. This is achieved by re-sampling images in the latent space of the U-Net to a standard resolution and then restoring the original resolution prior to output.

Denoising Diffusion Probabilistic Models (DDPM) [39] are methods for image generation, using an iterative denoising process learned by a neural network. The architecture is flexible but usually have a U-Net or Transformer [40] component. The core benefit of the denoising process is preventing regression to the mean during learning and inference. Instead, the model produces a random walk through the space of possible data.

2.2.2. Robust estimators in registration

A straightforward extremum estimator is the **Least-Squares** (LS) estimator. It minimizes the sum of squared residuals $\sum_i r_i^2$, where r_i is the i -th residual. This standard operator is highly sensitive to outliers, as squaring large residuals assigns them a disproportionately large weight.

Tukey's biweight function [41], on the other hand, is a robust criterion function. It is often used for optimization in registration tasks due to its “redescending” property (see Figure 2.3). Here strong outliers are completely ignored, providing high robustness. The function is defined as:

$$\rho(r) = \begin{cases} \frac{c^2}{6} \left(1 - [1 - (r/c)^2]^3\right) & \text{if } |r| \leq c \\ \frac{c^2}{6} & \text{if } |r| > c \end{cases} \quad (2.3)$$

The **Huber function** [42] behaves like the standard LS for small residuals but switches to the robust L_1 norm for residuals larger than a threshold c . This maintains high efficiency for small errors while reducing, but not discarding, outliers. It is formulated as:

$$\rho(r) = \begin{cases} r^2/2 & \text{if } |r| \leq c \\ c|r| - c^2/2 & \text{if } |r| > c \end{cases} \quad (2.4)$$

The **robust iterative closest point algorithm** (robust ICP) [43] incorporates a robust extremum estimator into the standard ICP algorithm. Instead of finding the transformation that minimizes the sum of squared distances (LS) between matched points

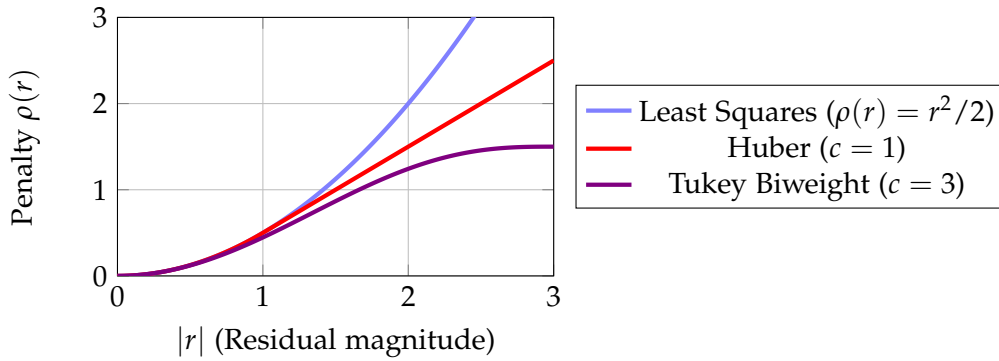


Figure 2.3.: The behavior of criterion functions Least Squares, Huber, and Tukey's Biweight with increasing residuals. The thresholds c are chosen to best illustrate criterion differences.

like the standard ICP [44], it minimizes a robust criterion in every iteration. We use the **iterative reweighted least squares** to solve the non-linear criteria function $\sum \rho(r_i)$ efficiently. This method linearizes the problem by reformulating the criterion function to a weighted least-squares (WLS) problem. In particular, the objective function is rewritten as $E_{WLS} = \sum_i w_i \cdot r_i^2(\mathbf{T})$, where $\mathbf{T}$ is the transformation and w_i is the weight given to the i -th residual. To find the weight w_i , use the formula $w_i = \frac{\psi(r_i)}{r_i}$, where $\psi(r) = \rho'(r)$ is the influence function (see [45]), and r_i is the residual from the prior iteration. This formulation is not only efficient but also allows for straightforward visualization of weights and therefore influence of outliers, as well as easy intervention during registration with additional modifications of w_i .

2.2.3. Evaluation metrics

There are multiple avenues for evaluating method performance for medical image analysis. One option is to perform end-to-end analysis on a population with a known analysis outcome (as described in the previous chapter). If the method finds the expected significant differences, we can consider the method able to detect the investigated morphological changes. This approach, however, covers only one population per test and only evaluates performance on already known cases. For more interpretable analysis on heterogenous data, we compare our methods' (intermediate) outputs to a reference standard. This data can be manually or synthetically generated. Depending on the type of outputs we employ a set of difference metrics, that are introduced below.

Segmentation

The **Dice Similarity Coefficient** (DSC) [46, 47] measures the overlap of two binary segmentation masks X, Y :

$$\text{DSC}(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|}, \quad (2.5)$$

where 1 describes a perfect match and zero indicates no overlap between the masks.

The **Hausdorff Distance** (HD) measures the maximum distance from a segmentation boundary in one mask X to the boundary points in the other set Y [48]:

$$\text{HD}(X, Y) = \max \left\{ \sup_{x \in X} d(x, Y), \sup_{y \in Y} d(X, y) \right\} \quad (2.6)$$

For more robust estimates we use the 95th percentile (HD95).

Registration

The **Head Pose Difference** (HPD), also known as **Jenkinson's RMS Deviation** [49] quantifies the difference between two rigid or affine transformations T_1, T_2 , by first splitting the difference $T_2 T_1^{-1} - I$ into rotational component A and translational component t . Then the difference in millimeters can be estimated as an integral over a sphere with the size of an average head:

$$\text{HPD}(A, t) = \sqrt{\frac{1}{5} R^2 \text{Trace}(A^\top A) + (t + Ax_c)^\top (t + Ax_c)}. \quad (2.7)$$

For more details, please refer to the original tech report [49].

Image generation

The **Structural Similarity Index Measure** (SSIM) is a perceptual metric quantifying image quality degradation by comparing local patterns of pixel intensities normalized for luminance and contrast [50]. The SSIM between two images x and y is defined as

$$\text{SSIM}(x, y) = \frac{(2\mu_x \mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}. \quad (2.8)$$

An SSIM of 1 is a perfect match of x and y and -1 is the lowest possible score. Similarly, the **Peak Signal-to-Noise Ratio** (PSNR) is the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the image fidelity [50]:

$$\text{PSNR}(x, y) = 10 \cdot \log_{10} \left(\frac{M^2}{\text{MSE}(x, y)} \right). \quad (2.9)$$

The PSNR measures image similarity in decibel (dB). Both metrics capture slightly different aspects of image similarity. Due to the complex nature of the comparison, reporting multiple metrics is preferable.

Signal processing

To measure the similarity of two variables (e.g., time series), despite different mean and variance, we can use **Mutual Information** (MI):

$$MI(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right) \quad (2.10)$$

2.2.4. Statistical testing

In addition to statistical medical image analysis (described in the previous section), we use statistics to test whether our proposed methods significantly outperform state-of-the-art methods.

The **Wilcoxon Signed-Rank Test** [51] is a non-parametric paired difference test, suitable to assess whether higher metrics (e.g., DSC) of one method are significantly higher than another. Importantly, it does not assume a Gaussian distribution in the input sampled (like, for example, Student’s t-test). This is important, as variation in medical imaging and method effects can cause shifted distributions and strong outliers (e.g., in the case of motion artifacts or disease).

For categorical data in small datasets, **Fisher’s exact test** [52] is used to determine method differences. This test is also non-parametric and useful for categorical manual annotations of method results (e.g., better, worse, equal). When data sizes are larger, the **chi-squared test** [53] is preferred.

For comparing the similarity of measurements, instead of their differences, we use the **Intraclass Correlation Coefficient** (ICC) [54], specifically measuring the degree of absolute agreement among measurements (criterion-referenced reliability [55]), which compares equality, not only correlation. This is especially useful to establish test-retest reliability of our methods, i.e., whether two MRI scans of the same person yield the same measurements, despite potential variations in head pose and scan settings and quality.

Chapter 3.

Quantifying MR head motion in the Rhineland Study – A robust method for population cohorts

This chapter is based on the peer-reviewed publication by Pollak et al. [2], which addresses error and bias caused by head motion at the first and the final step of the previously described analysis process. In this paper, we propose a robust registration method that can measure head movement during MRI acquisition based on depth images taken by an optical camera. These measurements can be used in statistical analysis to correct the systematic errors that are caused by head motion [56] or to further study motion effects.

Quantifying MR head motion in the Rhineland Study – A robust method for population cohorts.

Authors: **Clemens Pollak**, David Kügler, Monique M. B. Breteler, and Martin Reuter.

Published in: NeuroImage, 275, 120176.

DOI: [10.1016/j.neuroimage.2023.120176](https://doi.org/10.1016/j.neuroimage.2023.120176)

Github: [Deep-MI/head-motion-tools](https://github.com/Deep-MI/head-motion-tools)

3.1. Summary

3.1.1. Motivation

Statistical analyses of MRI in large cohort studies and clinical trials generally consider confounding variables and demographic factors. However, they frequently overlook potential confounders associated with MR acquisition, such as discrepancies in sequence, hardware, or firmware. A prominent confounder of MRI analysis is head motion. Head motion during the MRI acquisition has been demonstrated to systematically distort MRI-based neuromorphometric analysis [19, 57–64]. Motion-induced imaging artifacts can lead to erroneous conclusions regarding diminished gray matter volume estimates in T1-weighted MRI analysis [19, 60–64], and even persist despite expert quality control [19, 61]. These motion-induced biases are anticipated to increase with the implementation of higher-resolution imaging sequences [65] and prolonged scan durations per sequence. Furthermore, increased head motion is frequently associated with research variables such as age [62, 66], body mass index (BMI) [67, 68], and various neurological disorders, including autism [69–71], Parkinson’s [72], Alzheimer’s [73], and Huntington’s disease [74], which may lead to an under- or overestimation of these effects.

3.1.2. Contributions

To address the aforementioned challenges, we propose to use an iterative closest point registration algorithm with a robust criterion function and an additional stabilizer to dampen movements at the back of the head. We also introduce an aggregated *motion score* for precise head tracking and motion quantification during MR acquisition using a commercially available depth camera. Our registration produces a time series of head poses (*motion trace*) by aligning individual point cloud data obtained from depth video frames to a reference frame. Our approach significantly outperforms the vendor-provided head tracking technique [11, 75] across three novel assessments: 1. increased similarity with functional MRI motion estimates, 2. greater mutual information with high-frequency respiratory measurements, and 3. superior correlation of our motion score with multiple image quality metrics. Additionally, we reproduce known correlations of motion with age and BMI using the proposed method. Ultimately, we implement our method in the Rhineland Study [76, 77], marking the inaugural application of optical tracking in a large population cohort study. Unlike other studies that focus on datasets with high motion levels, our approach tackles the head tracking task within a general population cohort marked by low motion levels, requiring heightened sensitivity and the ability to track hundreds of participants with minimal

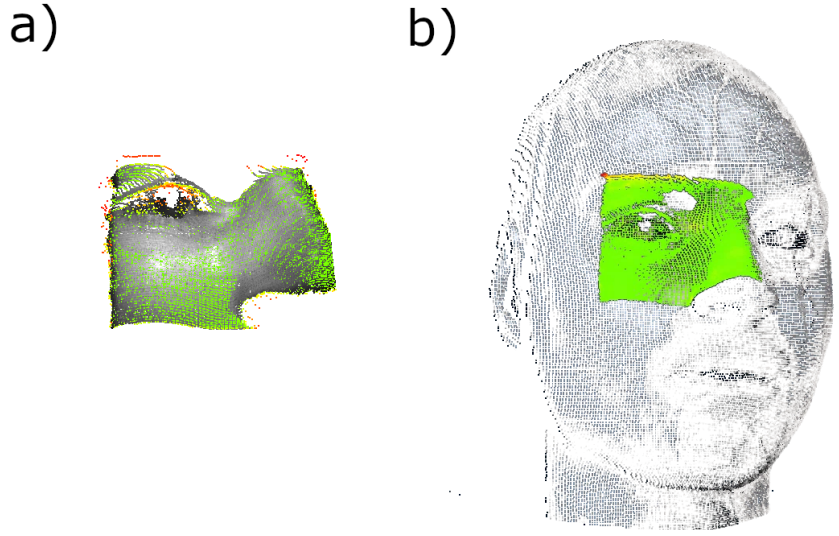


Figure 3.1.: Robust alignment of motion capture (green to red) with a) a reference point cloud and b) structural MRI (grey to black). Red and yellow areas are assigned lower weights during robust registration.

intervention. In the Rhineland Study, we further validate our method by confirming previously known correlations of BMI and age with motion. We also find that head motion increases over the course of the scan session and that motion measured from functional MRI (fMRI) acquisitions can serve as a reasonable proxy of motion in adjacent scans. fMRI-derived head motion can therefore be used as a surrogate control variable when no dedicated head tracking is available.

3.1.3. Motion tracking pipeline

Up to seven distinct MRI sequences are obtained by the scanner during the one-hour MRI session in the Rhineland Study protocol [78]. Throughout the session, simultaneous optical head tracking and respiration tracking are carried out. Head tracking is facilitated by a depth camera (TracInnovations, Copenhagen, Denmark) [11, 79, 80] mounted on the head coil of the scanner. It records a portion of the participant's face during the scan. The captured facial regions are acquired at different intervals, contain artifacts, and vary based on the head's size and position, making motion tracking challenging. To address this, we propose a modified version of the iterative closest point algorithm [81, 82], where we choose Huber's function as the most suitable criterion function and introduce an anatomically motivated stabilizer, which reduces

motion at the back of the head. Using this registration method, we compute a motion trace by rigidly registering all point clouds (generated from depth images) to the manually cleaned reference point cloud (see Figure 3.1 a), providing information on how much the participant’s head moved during the scan.

After acquiring a series of head poses via registration, we correct for synchronization errors between the motion camera and the MRI scanner using a novel metric called *motion trace difference*, which quantifies the transformation distance between two time series in an unbiased fashion. For synchronization, we minimize the difference between scanner-based fMRI motion estimates and motion camera-based estimates by changing the time difference between MRI and motion camera measurements. Next, we ensure that the motion estimates are equidistant by resampling the translation vectors and quaternions [83] of the motion trace, using a sliding window with a triangular weight function for averaging. The correctly time-stamped, equidistant head poses can then be summarized in a motion score for each acquired MR image. We propose to use the average head velocity during the scan and calculate it by calculating the *head pose difference* (HPD) between consecutive measurements. HPD is derived by averaging the displacements within a ball, which represents the head [49]. These displacements reflect the movement from one pose to another (described by transformation matrices), essentially weighting translational and rotational components of the transformation matrices.

3.1.4. Evaluation

Since this is the first publication that extensively validates head motion estimates derived from an optical in-MRI camera in a large population cohort, we propose several novel evaluation criteria. These criteria use correlates and other estimates of head motion to validate our method, since no gold-standard method exists. We propose four evaluations:

1. **Difference to fMRI-based motion estimates:** We extend the HPD to compare series of head poses (*motion trace difference*, MTD) instead of single time points by applying our re-sampling and registration to T1w images (see Figure 3.1 b) and then comparing head poses independent of their reference by considering every time point as a reference starting point. MTD provides a fair and symmetric difference measure between our method and reference motion estimates derived from an established fMRI analysis toolbox (McFLIRT [84]).
2. **Mutual information with respiratory measurements:** We utilize respiration information captured during the scan to validate that we can detect respiration in the head motion estimates derived with our method. For this we sample both

signals at the same time point, then keep the translational and rotational parts as separate values to calculate the mutual information with each value to the respiratory measurements.

3. **Correlation with image quality metrics:** To see whether the motion estimates derived with our method show motion relevant to the acquired MRI's quality, we use the MRIQC [85] toolbox to extract five quality metrics for T1w MR images. Then we test the correlation of our motion score for the T1w acquisition with the quality metrics.
4. **Correlation with BMI and age:** BMI and age are two well-known correlates of head motion. We reproduce these known correlations in the Rhineland Study using motion scores derived by our method.

Overall this evaluation scheme uses accurate low-frequency measurements (fMRI), relative high-frequency measurements (respiration), image quality, and research variables to test different aspects of tracking methods, giving reliable assessments of accuracy in the absence of ground truth measurements.

3.1.5. Results

We use the proposed evaluations to compare our method to the vendor-supplied tracking method and a baseline (dummy method). The proposed motion tracking method outperforms the vendor-supplied method and baseline methods in all evaluations, indicating superior robustness and accuracy. Additionally, since the optical tracking method is independent of the scanner, we can, for the first time, directly compare motion in a one-hour scan session across multiple MR sequences. We observe that motion increases consistently over time (approximately 29% after 30 minutes of scan time). This means that the sequences most affected by head motion should be scheduled first, and the order of sequences in a protocol should remain the same throughout a study. Furthermore, we find that in the absence of a motion camera, motion estimates from fMRI can be used to approximate motion during later scans.

3.1.6. Conclusion

In summary, we introduced a fully automatic, extensively validated motion-tracking pipeline. The tracking tools are available in an open-source repository (github.com/DeepMI/head-motion-tools). The optimized registration algorithms allow users to register point clouds in real time. The tool is currently used in the Rhineland Study, a large population cohort with over 10,000 participants. Other studies with optical head tracking can equally benefit from the tracking and evaluation tools.

3.2. Author Contributions

In this publication, Clemens Pollak designed the project in collaboration with David Kügler and Martin Reuter. He further developed and implemented all stages of the motion tracking pipeline and conducted experiments to validate it. He curated motion camera data, which was recorded and provided by the Rhineland Study team led by Monique Breteler, along with curated MRI. Clemens analyzed the motion effects and method performance. Together with David Kügler and Martin Reuter, Clemens wrote the first and final drafts, including reviewing and editing, with valuable input from Monique Breteler. Finally, Clemens deployed the open-source tool.

Chapter 4.

Estimating head motion from MR-images

This chapter is based on the peer-reviewed publication by Pollak et al. [3]. It expands on the previous chapter by predicting a motion score directly from the MRI without the need for a motion camera.

Estimating head motion from MR-images.

Authors: **Clemens Pollak**, David K  gler, and Martin Reuter.

Published in: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), pp. 1-5. IEEE

DOI: [10.1109/ISBI53787.2023.10230717](https://doi.org/10.1109/ISBI53787.2023.10230717)

Github: [Deep-MI/head-motion-from-MRI](https://github.com/Deep-MI/head-motion-from-MRI)

4.1. Summary

4.1.1. Motivation

As established in the previous chapter, head motion is a ubiquitous challenge for magnetic resonance image (MRI) acquisition. It causes a range of image artifacts that introduce bias in downstream analysis [19, 60–64], which persists despite expert quality control [19, 61]. Therefore, prior tools that predict expert motion ratings [86–94] cannot capture all relevant artifacts. This is especially relevant when considering compliant, healthy participants of the large population studies like the Rhineland Study, where few cases are found to contain artifacts upon visual inspection. The development of our motion tracking toolbox now enabled us, for the first time, to estimate an objective motion score directly from MRI sequences.

4.1.2. Contributions

We proposed a tool that can predict motion scores directly from T1-weighted, T2-weighted, and FLAIR MR images. For this purpose we present a deep-learning method, which outperforms DenseNet and state-of-the-art quality/motion estimation methods on the challenging Rhineland Study dataset. Furthermore, we show that our learning setup can even detect distinct artifacts of motion from respiration and (relaxation) drift. Finally, our method detects the significant, known correlation between predicted motion and age.

4.1.3. Motion estimation pipeline

In our motion prediction pipeline we choose to avoid standardization, since subtle motion artifacts could be masked by interpolation artifacts. One exception is intensity normalization—interestingly, here we find that reducing the image to the 8 least significant bits in integer representation enhances prediction accuracy, potentially focusing the network on fine image differences (see Figure 4.1 left). Then, for prediction of motion scores, we adopt a fully 3D convolutional neural network (CNN) architecture [37]. Instead of directly regressing the motion score, we follow the scheme proposed by Peng et al. [37] for brain age prediction and use a classifier-style neural network. For this approach we generously define the expected range of motion and split this range into 40 bins of "prototypes". The network then predicts the probability that the current motion score belongs to the prototype. For training we use the Kullback-Leibler loss. To reconstruct the motion score from the predicted probability distribution, we sum the product of prototype centers and predictions.

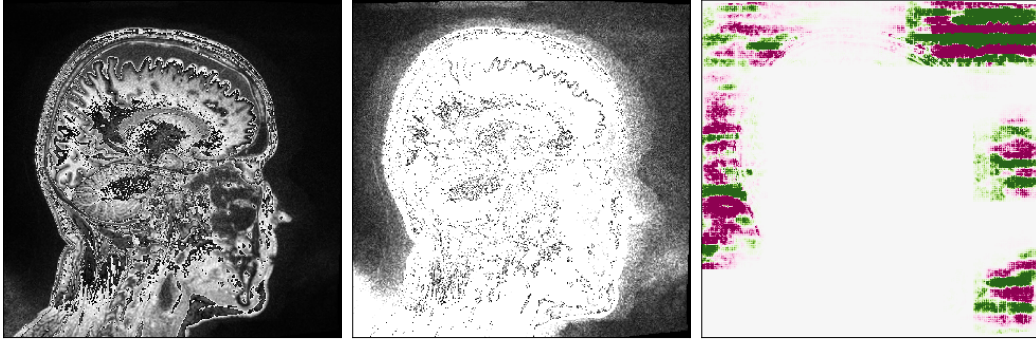


Figure 4.1.: Interpretability analysis for our motion detection model. On the left is the preprocessed input MRI; in the middle the intensities are scaled to focus on the background for reference. On the right we see the results of gradient-weighted class activation mapping (GRAD-CAM), which seems to show background artifacts highlighted. Areas highlighted in green support the predicted class (positive influence), while red areas support other classes (negative influence), white areas have no strong impact (values close to zero).

4.1.4. Evaluation

First we validate our choice of architecture by comparing our method to i) the classical DenseNet [95], ii) the model our architecture is based on, SFCN [37], and iii) a CNN specifically designed for quality estimation from MRI (MIQA) [93]. We use the coefficient of determination (R^2) and Spearman's ρ to test agreement with the camera-derived scores. Next, we use the same evaluation to compare our method to existing tools that can estimate head motion from images. First we test four metrics for motion/quality from the MIQA tool [93], then we test the average edge strength (AES) [96], which is a known predictor of head motion during structural MRI [96]. For final validation, we verify that our method can detect the same correlations with BMI and age that we could detect with the camera-derived motion estimates. To showcase the generalization of our method, we also train it on T2-weighted and FLAIR MR images, where no method for motion detection exists currently. Finally, we also test whether our method can be retrained to detect different kinds of movement, like respiration and slow drifting motion that occurs when neck muscles relax throughout the scan session.

4.1.5. Results

Our method outperforms the other retrained architectures and state-of-the-art methods across the board and can approximate motion camera estimates. Our method achieves good scores on three different MRI modalities, showing that the method is not specific to a protocol or acquisition scheme and can be used where currently no other tools exist. In further unpublished interpretability analysis (see Figure 2.2) we see that the network indeed seems to focus on motion artifacts with high contrast, typically in the image background. It is even able to detect the magnitude of respiration or slow relaxation drift, two movements that cause fundamentally different artifacts in the scanner. Furthermore, we find significant correlations with BMI and age as expected. This demonstrates that our method is a suitable replacement for measuring head motion during the MRI acquisition.

4.1.6. Conclusion

We introduce the first method to predict objective head motion estimates from MR images. Our method learns the relationship between the complex MRI acquisition pipeline and the large variety of head movements during the scan and can accurately predict motion during the scan from an image alone. Furthermore, during application in the Rhineland Study, we were able to observe that our method perfectly separated scans that contained visible artifacts from scans without visible artifacts. This showcases that it is not only a tool for motion detection but also a valuable quality control tool. A limitation of the current method is that the learned relationship between MRI acquisition and motion artifacts is highly scanner and protocol specific – therefore, retraining is likely necessary for unseen scanners and acquisition protocols. To enable retraining our method and use of our tools with the existing protocols, we release it in an open-source repository (github.com/Deep-MI/head-motion-from-MRI).

4.2. Author Contributions

In this publication Clemens Pollak designed the project in collaboration with David Kügler, and Martin Reuter. He further developed and implemented all stages of the motion estimation method and conducted experiments to validate it. He curated motion camera data, which was provided by the Rhineland Study team lead by Monique Breteler, along with curated MR images. Together with David Kügler and Martin Reuter, Clemens wrote the first and final drafts, including reviewing and editing. Finally, Clemens deployed the open-source tool.

Chapter 5.

FastSurfer-LIT: Lesion inpainting tool for whole-brain MRI segmentation with tumors, cavities, and abnormalities

This chapter is based on the peer-reviewed publication by Pollak et al. [4] and presents a novel method for processing structural MRI with damaged brain tissue (brain lesions). It leverages Diffusion Denoising Probabilistic Models (DDPM) to replace damaged tissue with healthy tissue. This enables tools that originally are not designed to handle MRI with damaged tissue, like the established FreeSurfer and FastSurfer toolboxes, to aid in the analyses of intervention effects, e.g., side effects of surgery or radiotherapy treatment.

FastSurfer-LIT: Lesion inpainting tool for whole-brain MRI segmentation with tumors, cavities, and abnormalities.

Authors: **Clemens Pollak**, David Kügler, Tobias Bauer, Theodor Rüber, and Martin Reuter.

Published in: Imaging Neuroscience (2025), 3.

DOI: [10.1162/imag_a_00446](https://doi.org/10.1162/imag_a_00446)

Github: [Deep-MI/LIT](https://github.com/Deep-MI/LIT)

5.1. Summary

5.1.1. Motivation

Tumors, resection cavities, and other lesions can fundamentally alter brain structure and present as abnormalities in brain MRI. In some cases neuroimaging pipelines like *FreeSurfer* [23], *FSL* [97], or *SPM* [98] can generate results for images with lesions [99]. None of these tools, however, are developed for or validated on images with lesions. When the present lesions are large, traditional pipelines and recent deep learning tools frequently fail catastrophically by either crashing or producing obviously flawed outputs. Even when standard toolboxes completed without obvious failures, the measurements on images containing lesions were found to be biased [100]. Meanwhile, the study of disease effects and the study of intervention (side) effects require analyses on brain MRI with pathologies. To address this gap, researchers require specialized tools, which can generate accurate morphometric measurements in the presence of such pathologies and cavities.

5.1.2. Contributions

We introduced *FastSurfer-LIT*, a pipeline designed to perform whole-brain segmentation and surface reconstruction in the presence of small and large lesions, as well as cavities – while operating across various resolutions, scanners, and lesion types. The lesion inpainting (LIT) network is the pipeline’s first stage. We propose a novel volumetric and resolution-independent Denoising Diffusion Probabilistic Model (DDPM) to enable high-quality inpainting for arbitrary lesion shapes at multiple desired resolutions. This network architecture can operate on multiple resolutions with high accuracy and allows inpainting for any size or shape of mask. Additionally, we present an inference scheme that uses the reverse diffusion process to transmit spatial information. This enables us to present the first structural MRI inpainting technique that natively operates on sub-millimeter MRI while maintaining low memory requirements. We integrated the lesion inpainting tool LIT with *FastSurfer* (for whole brain segmentation and surface reconstruction) to create the *FastSurfer-LIT* neuroimaging pipeline. We also added post-processing tools to the pipeline to remove abnormal regions from volumetric and surface statistical analysis, allowing for seamless group analysis that includes participants with brain lesions.

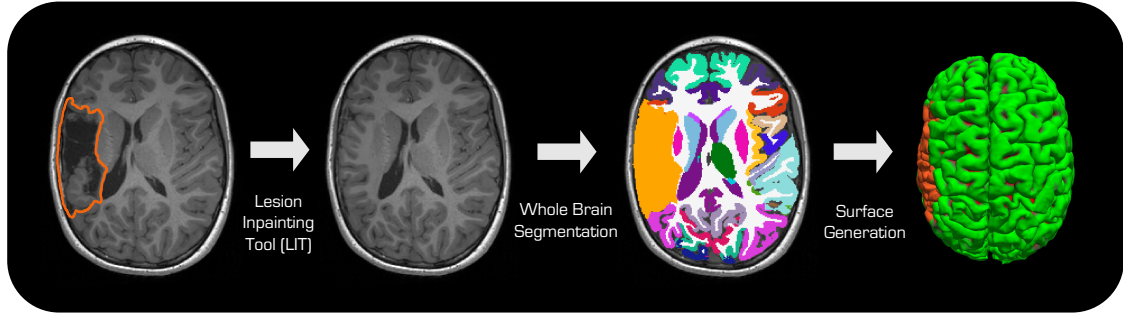


Figure 5.1.: Overview of the proposed lesion-robust whole-brain analysis pipeline.

The Figure is courtesy of Imaging Neuroscience, authors Pollak et. al [4], CC-BY 4.0

5.1.3. Lesion-robust whole brain analysis pipeline

The proposed whole brain analysis pipeline for MRI with lesions requires a structural MRI and a lesion mask, which marks the area of damaged tissue (see Figure 5.1). The pipeline can be divided into three distinct steps:

1. **Lesion inpainting:** The first and central step of the pipeline is the lesion inpainting tool. For the inpainting task, we propose a novel, resolution-independent [24] unconditional DDPM, which is trained unsupervised on a lesion-free, structural MRI. Then we modified the DDPM inference by keeping the area outside of the lesion mask fixed (but with an appropriate noise level), essentially guiding the network to generate an anatomically plausible healthy patch of tissue. Additionally, we expand the memory-intensive 2D U-Net architecture to volumetric data by training three view-specific 2D networks (for axial, coronal, and sagittal MRI slices). Then we apply the networks in an alternating fashion, which generates spatially consistent 3D images with minimal additional computational cost. We combine this technique with multi-channel 2D networks (2.5D). These networks process a rotating set of image slabs, which further disseminates spatial information and image generation artifacts.
2. **Whole brain segmentation:** For whole brain segmentation we use the pre-trained FastSurfer whole brain segmentation network [21, 24]. To ensure high segmentation quality, our in-painting method is trained on the same data as the segmentation network, which means it generates data that is close to the training distribution. After segmentation, resulting statistics from the whole brain segmentation are adjusted for the lesion and exclude inpainted areas.
3. **Cortical surface reconstruction:** Similar to the previous step, we use FastSurfer’s cortical surface reconstruction pipeline. Then, after surfaces are reconstructed, we

map the lesion mask onto the cortical surface maps. These surface labels are then used to exclude lesion-affected areas from analysis of cortical folds or thickness. Since these regions can be excluded on a participant level, group-level statistics can still be generated for a complete cortical surface model.

5.1.4. Evaluation

We compare our tool with two recent methods capable of lesion inpainting: SynthSR [29] and VBG (Virtual Brain Grafting) [99]. To demonstrate that our approach is a general standardization approach to enable processing of lesion-affected MRI, we do not only use FastSurfer for segmentation and surface reconstruction but also replace it with FreeSurfer in all methods and repeat the evaluations. For direct comparison, we generate two evaluation datasets: i) A dataset of glioblastoma computationally grown in healthy brain MRI [101] and ii) multiple sclerosis lesions transferred from patients onto healthy MRI via registration. This procedure provides us with pairs of healthy and artificially diseased images, where we can use the segmentations of healthy brains as a reference standard. During data processing, we also track runtimes, crashes, and surface reconstruction errors. Using real clinical samples, we evaluate the test-retest reliability of cortical surfaces before and after temporo-mesial resection surgery and also perform a blinded comparison study with experts rating the segmentation quality based on different inpainting methods. This clinical dataset is highly heterogenous and includes 14 different types of brain lesions.

5.1.5. Results

Our method significantly outperforms others, using either FreeSurfer or FastSurfer as segmentation module. Notably, this is true for medium to large lesions (glioblastoma) and small lesions (multiple sclerosis) with very different distribution patterns. Our method also outperforms the state-of-the-art (VBG) in cortical surface reconstruction reliability. One likely reason for this is that our method uniquely ensures that the areas outside the lesion mask remain unaltered during the inpainting process. Finally, we observe no crashes or other failures when running our method and see that it is much faster than the state-of-the-art (roughly a two-hour speedup on a desktop machine).

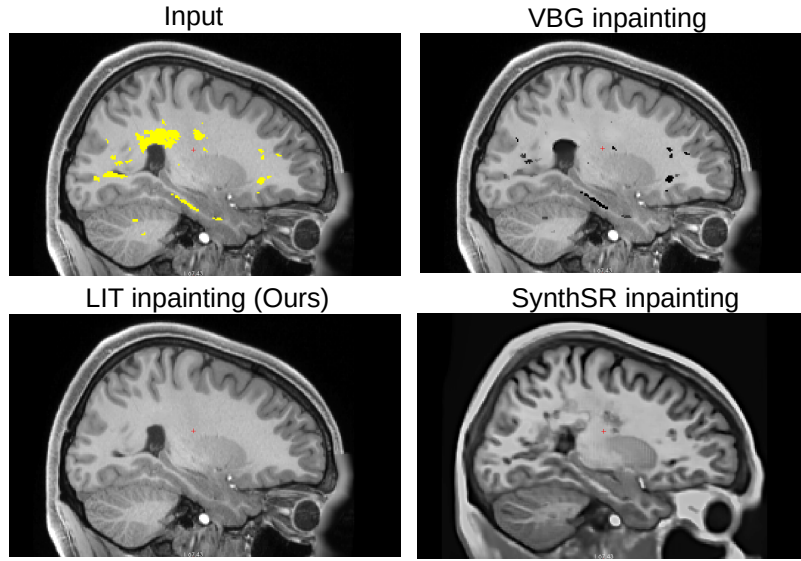


Figure 5.2.: Qualitative comparison of inpainting results for a case with simulated multiple sclerosis lesions.

5.1.6. Conclusion

In this work, we introduce FastSurfer-LIT, a whole-brain segmentation and surface reconstruction pipeline for structural MRI with lesions and cavities. Besides leaving areas outside of the specified lesion mask unchanged, our learned brain model also allows LIT to generate tissue that is completely individualized to the target brain, leading to more plausible inpaintings compared to template-based approaches. With these contributions, we hope to accelerate research on the impact of interventions and disease (e.g., radiotherapy, surgery) on overall brain health. FastSurfer-LIT could grant new insights into cortical reorganization [102], with higher resolution data, more accurate cortical reconstruction, and the ability to exclude only the lesion area, instead of limiting the analysis to areas where all patients are lesion-free. Besides accelerating and improving research, this work can also contribute to fairness and accessibility of personalized medicine (e.g., personalized structural connectomics [103]), which can be made available for individuals with abnormal brain structures. Since LIT can run independently from the downstream segmentation or surface reconstruction steps of the pipeline, it can easily be combined with other neuroimaging tools. In our evaluation, we focus on segmentation and surface reconstruction with challenging data. We expect LIT to also enable use of other neuroimage analysis software with lesions, such as structural sub-segmentation [104, 105] or subject-to-subject and subject-to-atlas registration (previously demonstrated by SynthSR [29, 106]). Therefore, we release both

a standalone version of the tool (github.com/Deep-MI/LIT) and an integration into the FastSurfer toolbox (github.com/Deep-MI/FastSurfer).

5.2. Author Contributions

In this publication Clemens Pollak designed the project in collaboration with David Kügler, and Martin Reuter. He further developed and implemented all stages of the lesion inpainting method and the FastSurfer-LIT pipeline and conducted experiments to validate them. Tobias Bauer and Theodor Rüber provided the data for evaluation from the Center for Epileptology at the university clinic in Bonn, along with valuable insight regarding the clinical application of the tool. Together with David Kügler and Martin Reuter, Clemens wrote the first and final drafts, including reviewing and editing. Finally, Clemens deployed the open-source tool.

Chapter 6.

FastSurfer-CC: A robust, accurate and comprehensive framework for corpus callosum morphometry

This chapter is based on the research paper by Pollak et al. [5] currently under review at Imaging Neuroscience. The work introduces a comprehensive corpus callosum morphometry framework that includes localization of landmarks, rapid standardization of the head pose, segmentation, surface reconstruction, and shape analysis. FastSurfer-CC outperforms existing specialized tools across the individual tasks and reveals statistically significant differences between Huntington’s disease patients and healthy controls that are not detected by the current state-of-the-art.

FastSurfer-CC: A robust, accurate, and comprehensive framework for corpus callosum morphometry.

Authors: **Clemens Pollak**, Kersten Diers, Santiago Estrada, David Kügler, and Martin Reuter.

Published in: arXiv, under review

DOI: [arXiv.2511.16471](https://arxiv.org/abs/2511.16471)

Github: [Deep-MI/FastSurfer](https://github.com/Deep-MI/FastSurfer)

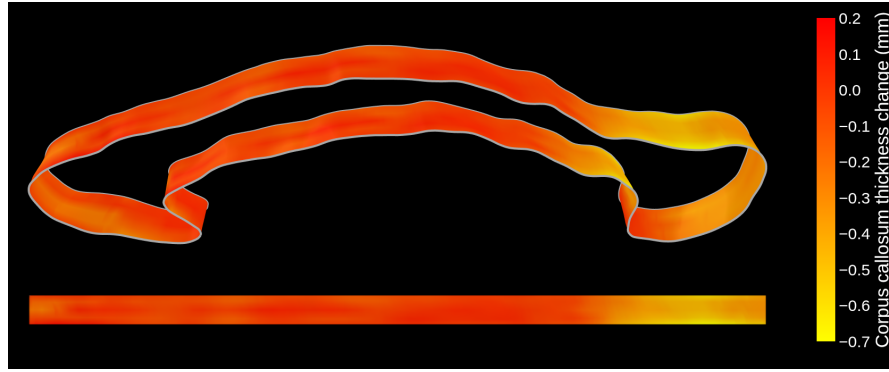


Figure 6.1.: Change of corpus callosum thickness between a group and controls and Huntington’s disease patients, displayed on a template CC mesh.

6.0.1. Motivation

The corpus callosum is the largest commissure in the human brain. It connects the brain hemispheres, which makes it a central focus in research on aging and neurological diseases, like epilepsy [107], autism [108], schizophrenia [109], multiple sclerosis [110], cerebral palsy [111], Parkinson’s disease [112], bipolar disorder [113], and Alzheimer’s disease [114]. Additionally, it is a target for therapies like deep brain stimulation [115] and is a central biomarker in clinical trials on remyelination therapies [116]. Even though just the size of the corpus callosum serves as a valuable biomarker, atrophy in particular regions correlates with various diseases and is of extensive research interest. Consequently, local sub-segment areas and thickness variations throughout the corpus callosum are significant morphometric factors. Although corpus callosum segmentation has been studied extensively, there aren’t many publicly accessible tools that offer a comprehensive, automated analysis pipeline. We identify five steps that are required for robust and accurate CC analysis:

1. **Find the mid-sagittal plane:** The lack of clear CC boundaries in the lateral direction motivates finding an accurate mid-sagittal plane and is often addressed by segmenting the CC in mid-sagittal slices of the brain within a pre-defined width [23, 117, 118]. Accurately determining mid-sagittal slices independently of the image slicing direction is crucial, as the shape, thickness, and area of the corpus callosum can vary significantly based on the position and orientation of the chosen mid-sagittal plane.
2. **Localize the anterior and posterior commissures:** The anterior and posterior commissures (AC & PC) are neuroanatomical landmarks used for localization of neuroanatomical structures during deep brain stimulation (DBS), stereotactic

and functional surgery [119], the definition of corpus callosum subsegments [120]. Their relative position is also a useful biomarker [121]. Here, they are required for multiple tasks: i) standardizing the head position in the remaining direction (nodding), ii) finding the corpus callosum end-points, and iii) estimating the CC location in the mid-sagittal slice as initialization for the segmentation task.

3. **Segment CC and fornix:** Segmenting the CC is at the core of most CC analysis pipelines. In midsagittal slices the CC is bright with high contrast to most surrounding structures. The fornix, however, frequently sits right below the CC and also appears as bright white matter in structural MRI. Previous tools [20, 117] struggled to differentiate the fornix and corpus callosum when they appear close to each other or touch. Therefore, we also include the fornix in this segmentation task.
4. **Derive morphometric estimates:** Advanced shape metrics can show changes in a more sensitive and interpretable way than simple area estimates. They also allow researchers to quantify local morphological changes, important to the understanding of aging and degenerative diseases. Therefore, deriving corpus callosum length, curvature, and thickness is a central part of the pipeline (illustrated in Figure 6.1).
5. **Subdivide the CC:** Finally, corpus callosum subdivision is a common task in neuroimage analysis, with a plethora of existing subdivision schemes for researchers to choose from. We observe, however, that individual variance in CC shape can introduce inconsistencies when performing subdivision with currently established schemes.

6.0.2. Contributions

Researchers who investigate corpus callosum changes in aging, disease, and intervention studies require an accurate tool for deriving markers of corpus callosum thickness, shape, area, and volume. Our framework, FastSurfer-CC, not only addresses all five sub-tasks in an open-source framework but also improves upon the existing approaches in every step. FastSurfer-CC:

1. Precisely determines the ideal midsagittal plane for corpus callosum examination.
2. Performs better than current specialized tools in segmenting the corpus callosum and fornix, as well as localizing the anterior and posterior commissures.
3. Demonstrates superior robustness and accuracy on challenging cases compared to other methods.
4. Presents new, reliable metrics for the length and curvature of the corpus callosum.
5. Provides efficient, precise, and robust thickness estimation along the corpus callosum.

6. Incorporates a novel shape-aware subdivision technique that works with existing schemes.

This integrated approach not only reduces processing inefficiencies but also ensures concerns of robustness and accuracy can be addressed with a holistic view. Overall, FastSurfer-CC is a fast, robust, accurate, and comprehensive tool for advanced corpus callosum morphometry.

6.0.3. Overview of the corpus callosum morphometry pipeline

Our pipeline is integrated with the FastSurfer whole-brain analysis toolbox. Therefore, we can use FastSurfer’s whole brain segmentation as a prior for estimating the mid-sagittal plane. Taking the centroids of all structures and registering them to a symmetrical template segmentation reliably aligns the head to an upright position, with the CC accurately divided along its left-right symmetry axis. This position also reveals the anterior and posterior commissures in mid-sagittal slices. To localize AC & PC, we use a 2D DenseNet [122], which is trained to predict coordinates. We use the Yuki [123] tool to generate a large corpus of pre-training data and then fine-tune the network with manual annotations. Next, we use the AC & PC locations to roughly estimate the CC location and use cropped mid-sagittal slabs to train a 2.5D VINN (Voxel-size Independent Neural Network [24]) U-Net to segment CC and fornix based on manual labels. Then we apply the marching squares algorithm to create contours and triangle meshes from the CC segmentations, essentially removing the further analysis from the image voxel grid. Central to our measurement approach is the CC midline, which is generated by solving the Laplace equation with a positive and a negative charge at the opposing inferior and superior boundaries. The length and curvature of the midline describe the length and curvature of the CC. We can also derive CC thickness estimates by rotating the Laplace gradients and measuring equally spaced level sets along the midline. This procedure provides a thickness profile with a user-defined resolution and allows comparison and tracking of local thickness changes. Finally, we also use the midline to introduce a novel sub-segmentation approach that cuts the CC into user-defined sub-segments. This geometry-informed approach is more robust to individual CC variance than previous approaches, which used a global direction for subdivision. In addition to our proposed pipeline, we also implement previously introduced and validated CC metrics (CC index [124] and CC circularity [125]), as well as five previously proposed subdivision schemes [23, 120, 126–128]. These additional features allow researchers to reproduce previous findings with our pipeline.

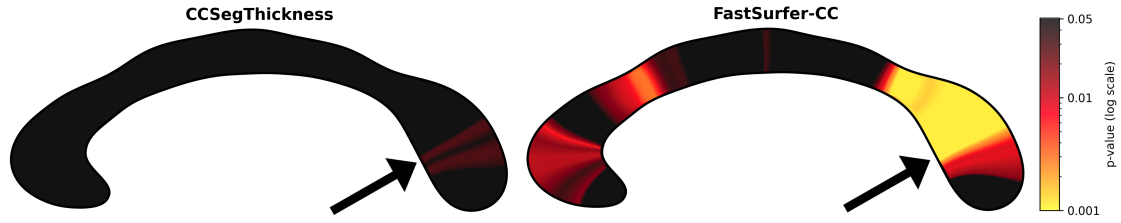


Figure 6.2.: Comparison of local CC thickness estimation for our method (FastSurfer-CC) and the previous state-of-the-art. Shown are significance (p-values) for a group comparison between Huntington’s disease patients and controls (adapted from [5]). While the methods agree in the area of highest significance (arrows), FastSurfer-CC finds more significant regions, with overall higher significance.

6.0.4. Evaluation

For the tasks of mid-sagittal plane detection, AC & PC localization, and CC & fornix segmentation tasks of our pipeline, we individually choose state-of-the-art methods for comparison and perform separate evaluations. Then we compare the sensitivity of our pipeline to the overall state-of-the-art in CC analysis with our method in a realistic research scenario. Following, we briefly list the task-specific evaluation methods. First, we compare the midsagittal planes generated by four methods in a blind manual comparison. For this, we use a dedicated rating tool, developed to compare mid-sagittal planes with respect to CC symmetry. Next, we compare our localization method to two other methods for localization of anterior and posterior commissures. We pay special attention to method robustness by using a randomly sampled test set and a more challenging test set (consisting of cases that were difficult to annotate manually). The same paradigm is also used for the evaluation of the segmentation network, which is compared to two dedicated CC segmentation methods (CCSeg [117], Yuki [118]), one method that also segments CC and fornix (FreeSurfer’s mri_cc [30]), as well as two fornix segmentation methods (SCLimbic [129], OSHy-X [130]). Finally, we use both CcSeg and our method to determine CC group differences between Huntington’s disease patients and controls.

6.0.5. Results

Our method outperforms the dedicated method for all evaluated sub-tasks. While the performance of some other methods drastically decreases for challenging cases, our method’s performance remains stable. Furthermore, our CC analysis pipeline is more

sensitive than the previous state-of-the-art CCSeg in the chosen group comparison. Our method reveals group differences missed by CCSeg in the comparison of Huntington’s disease patients and controls – both in aggregate metrics and local area and thickness changes (see Figure 6.2).

6.0.6. Conclusion

We propose a fast, robust, and accurate framework for corpus callosum morphometry. We have validated its components individually in a realistic group comparison scenario. Our evaluation is specifically designed to consider performance on challenging cases, which include brain lesions, strong atrophy, and artifacts, all of which were unseen during training. Furthermore, our method offers improved sub-segmentation approaches and additional downstream metrics, making it more comprehensive than previous tools. In addition to corpus callosum analysis, our framework facilitates head pose normalization, offering a rapid alternative to rigid template registration and serving as an initialization for other affine or non-linear registration tasks. The alignment requires mere seconds, enabling its use for preprocessing extensive datasets and its applicability in real-time applications. Our enhancements offer researchers a highly sensitive, innovative tool to investigate and assess changes in CC morphology related to disease and aging. FastSurfer-CC is available as part of the open-source FastSurfer toolbox (github.com/DeepMI/FastSurfer).

6.1. Author Contributions

In this publication Clemens Pollak designed the project in collaboration with David Kügler, and Martin Reuter. He further developed the corpus callosum analysis method, with support from Martin Reuter for shape analysis. Kersten Diers, Santiago Estrada, and David Kügler supported the project with data annotation and data curation. They also provided valuable insight and suggestions in the preparation of the manuscript. Together with David Kügler and Martin Reuter, Clemens wrote the first and final drafts, including reviewing and editing. Finally, Clemens deployed the open-source tool and incorporated it into the FastSurfer toolbox with support from David Kügler.

Chapter 7.

Summary and Outlook

This thesis presents three distinct approaches aimed at enhancing the robustness of neuroimaging pipelines through method innovation. The three main targets are i) reducing the negative effects of head motion, ii) brain lesions, and iii) measurement errors during CC analysis. We also scrutinize the robustness and generalizability of the methods we develop and thereby mitigate errors that are introduced by competing methods, like registration errors due to optical camera artifacts, slicing artifacts in DDPM generation, out-of-distribution errors in CC segmentation methods, and many more. Considering these additional sources of error is of utmost importance in method development for neuroimaging pipelines, as biomedical research depends on accurate and robust tools to deliver unbiased research results. To provide robust and accurate research tools, we have combined a large toolbox of methods, including deep learning (FCN, DenseNet, VINN, DDPM), registration (SVD, robust ICP, McFLIRT, etc.; HPD/RMSD), mesh-based shape analysis (Laplace, geometric), statistics (LME, ICC, R2, etc.), and visualization (point-cloud series, 2D/3D statistical analysis, segmentation, etc.).

With the proposed methods, we have also outlined general intervention points for the standard structural image analysis pipeline: image acquisition, standardization, and feature extraction. Improving the accuracy and robustness of feature extraction directly is often preferable to adding additional preprocessing steps because it does not introduce additional complexity and reduces the analysis runtime. In many cases, however (e.g., clinical cases with large lesions), such improvements are stifled by the absence of high-quality training data for deep learning methods. Additionally, good standardization approaches can improve a plethora of existing methods, while modifying feature extraction is usually done one method at a time. In other areas, like head motion, the effects on modern pipelines are currently not well researched and need to be established before developing more specific mitigation strategies. These examples showcase that choosing the point of intervention is a critical first step in neuroimage method development – and highly dependent on the error at hand. We

hope that the range of approaches in this thesis can provide multiple blueprints for approaching errors in neuroimaging pipelines from a method perspective.

All presented methods are available as open-source pipelines. Therefore, they offer direct access to more robust and accurate analysis for the research community. The motion-tracking tool is currently used [78] in the Rheinland Study, a large population cohort with over 10,000 participants. The lesion inpainting and corpus callosum analysis tools are integrated into the established FastSurfer toolbox for comprehensive structural neuroimage analysis. Following, we summarize the thesis contributions in more detail.

7.1. Head motion

The first major contribution of this thesis centers on the development of a robust framework for head motion quantification. We propose a registration method that leverages a commercially available depth camera to deliver precise head pose estimates during the MR acquisition. These estimates are then aggregated into a comprehensive motion score per image. Our method outperforms the vendor-supplied registration method, as evidenced by increased agreement with functional MRI-based motion estimates, higher mutual information with high-frequency respiratory traces, and superior correlation with a range of image quality metrics. Our method is used in the Rheinland Study, a large-scale population cohort characterized by low average motion, thereby necessitating high sensitivity and minimal intervention for effective tracking. This marks the first deployment of markerless motion tracking in a large cohort study. With the publication and open-sourcing of our novel evaluation criteria, in combination with the proposed tracking algorithm, we hope that markerless motion tracking technology can become more approachable for researchers in the community.

7.2. Whole-brain analysis in the presence of lesions

The second major contribution tackles the challenge of MRI analysis in the presence of structural abnormalities, like brain lesions and surgical cavities. We address this challenge by introducing FastSurfer-LIT, a pipeline for whole-brain segmentation and surface reconstruction that explicitly accommodates the presence of both small and large lesions, as well as complex anatomical variations, across a wide range of MRI resolutions and scanner types. At the core of the framework is a novel, volumetric, and resolution-independent denoising diffusion probabilistic model designed for high-fidelity lesion inpainting. This model enables accurate reconstruction of arbitrarily

shaped lesions at multiple resolutions. It considers information present in the 3D MRI space while keeping memory requirements comparable to 2D inpainting approaches. We show that the proposed inpainting leads to more accurate and more reliable segmentations and cortical reconstructions than previous state-of-the-art inpainting approaches – while also being faster. Beyond the presented application of inpainting for whole brain segmentation and cortical surface reconstruction, the method can also be applied to any other task currently inhibited by the presence of brain lesions or cavities.

7.3. Robust corpus callosum analysis

The third major contribution is the development and validation of FastSurfer-CC, a comprehensive open-source framework for corpus callosum morphometry. Unlike existing methods, FastSurfer-CC integrates all common sub-tasks into a unified, robust, and efficient pipeline, while also outperforming specialized tools. It i) accurately determines the optimal mid-sagittal plane for corpus callosum analysis, ii) outperforms specialized tools in the localization of the anterior and posterior commissures, as well as in the segmentation of the corpus callosum and fornix, iii) demonstrates superior robustness and accuracy in challenging anatomical cases, iv) introduces novel and reliable metrics for corpus callosum length and curvature, v) enables rapid and precise thickness estimation along the entire corpus callosum, and vi) implements a novel shape-aware subdivision approach that is compatible with previously established morphometric schemes. We specifically tested the method on unseen datasets using cases with lesions and cavities to assess the generalization of our deep-learning-based approach. In addition to rapid corpus callosum analysis within FastSurfer’s whole-brain segmentation pipeline, the tool also provides anatomically driven standardization of the head, serving as an effective preprocessing stage for subsequent analysis.

7.4. Future work

Our contributions open up research avenues for the development of new methods or improvements of existing tools. First, the proposed motion tracking tool can supply new insights into how study participants move in the scanner. This data is a valuable research target, as it can supplement existing motion studies on gait and tremor in the Rhineland Study with head motion during rest. Additionally, the scanner-external measurement of head motion can be applied to any MRI acquisition – allowing researchers to quantify how head motion affects the image quality and downstream estimates of fMRI, T1w,

T2w, FLAIR, and diffusion MRI sequences. A crucial challenge for this future work is mapping the head position during the acquisition to a single MRI measurement point. With a highly accurate synchronization and exchange of scanner information, we could predict the impact on image quality – which is, for example, higher when the head position differs a lot between low-frequency scanner measurements [26]. Even without accurate measurement of head motion, our modeling on quantifying head motion could provide a useful framework for quality estimation tools. Recent work [131], for example, is building on ideas proposed for our motion score and motion estimation network, but instead of using real motion traces, simulated motion is used to induce a large variety of artifacts that may be related to any acquisition sequence. With this scheme it may be possible to learn a measure of frequency space perturbation that generalizes across scanners and sequences.

Our proposed lesion inpainting tool also enables a plethora of new research avenues. With this tool researchers can now quantify the effects of radiotherapy treatment or brain surgery on the undamaged parts of the brain, empowering research on interventions with fewer side effects. Since the lesion in-painting tool is a general approach, this analysis is not limited to whole-brain segmentation and cortical surface reconstruction but could also be applied and validated with other pipelines, like voxel-based morphometry. Since the tool is completely independent of the area that shall be inpainted, it could also be used to replace other abnormalities, like local scanner artifacts. Re-training and expanding the tool to new modalities and cohorts is straightforward, since only raw MRI volumes are needed for the unsupervised training. A limitation of the in-painting approach in general is the need for a prior lesion segmentation. To generate lesion masks automatically, the method could be combined with an anomaly detection approach, some of which are also based on DDPMs trained on healthy data [132].

Finally, the corpus callosum morphometry tool could be adapted for the analysis of other tubular structures in the brain, like the fornix, the brainstem, ventricles, optical nerve, or arteries. Here we could similarly apply the proposed thickness, curvature, and length metrics with local components and apply localization and segmentation strategies for robust feature extraction. With our CC tool, we have shown that well-designed deep learning tools have remarkable generalization capabilities. To directly improve existing tools, we could incorporate our rapid, anatomically motivated head-pose standardization into existing toolboxes. Currently slow registration (e.g., with a Talairach template) is frequently required as an initial processing step. Since many of the template brains already follow the same anatomical alignment convention as our method, this procedure could be a drop-in replacement, potentially improving existing methods' speed and robustness.

7.5. Conclusion

Overall, this thesis proposes three distinct and comprehensive frameworks for addressing robustness concerns in neuroimaging pipelines. In each framework we introduce new methodological approaches touching the areas of registration, 3D time-series data, 3D diffusion denoising probabilistic models, segmentation, localization, robustness testing in neuroimaging, and robust shape analysis. The proposed tools will have a direct impact on the accuracy and robustness of brain analyses, as they are incorporated in the Rhineland Study or tightly integrated in FastSurfer, a widely used open-source toolbox for structural MRI analysis.

Appendix

Appendix A.

Publication: "Quantifying MR head motion in the Rhineland Study – A robust method for population cohorts"

This is an open-access article distributed under the terms of the Creative Commons 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

©2023 The Author(s). Published by Elsevier Inc.

Quantifying MR head motion in the Rhineland Study – A robust method for population cohorts.

Authors: **Clemens Pollak**, David Kügler, Monique M. B. Breteler, and Martin Reuter.

Published in: NeuroImage, 275, 120176.

DOI: [10.1016/j.neuroimage.2023.120176](https://doi.org/10.1016/j.neuroimage.2023.120176)

Github: [Deep-MI/head-motion-tools](https://github.com/Deep-MI/head-motion-tools)



Quantifying MR head motion in the Rhineland Study – A robust method for population cohorts

Clemens Pollak^a, David Kügler^a, Monique M.B. Breteler^{b,c}, Martin Reuter^{a,d,e,*}

^a AI in Medical Imaging, German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

^b Population Health Sciences, German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

^c Institute for Medical Biometry, Informatics and Epidemiology (IMBIE), Faculty of Medicine, University of Bonn, Bonn, Germany

^d A.A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston, MA, USA

^e Department of Radiology, Harvard Medical School, Boston, MA, USA

ARTICLE INFO

MSC:
0000
1111
MRI Head tracking
Population study
Depth camera

PACS:
0000
1111

ABSTRACT

Head motion during MR acquisition reduces image quality and has been shown to bias neuromorphometric analysis. The quantification of head motion, therefore, has both neuroscientific as well as clinical applications, for example, to control for motion in statistical analyses of brain morphology, or as a variable of interest in neurological studies. The accuracy of markerless optical head tracking, however, is largely unexplored. Furthermore, no quantitative analysis of head motion in a general, mostly healthy population cohort exists thus far. In this work, we present a robust registration method for the alignment of depth camera data that sensitively estimates even small head movements of compliant participants. Our method outperforms the vendor-supplied method in three validation experiments: 1. similarity to fMRI motion traces as a low-frequency reference, 2. recovery of the independently acquired breathing signal as a high-frequency reference, and 3. correlation with image-based quality metrics in structural T1-weighted MRI. In addition to the core algorithm, we establish an analysis pipeline that computes average motion scores per time interval or per sequence for inclusion in downstream analyses. We apply the pipeline in the Rhineland Study, a large population cohort study, where we replicate age and body mass index (BMI) as motion correlates and show that head motion significantly increases over the duration of the scan session. We observe weak, yet significant interactions between this within-session increase and age, BMI, and sex. High correlations between fMRI and camera-based motion scores of proceeding sequences further suggest that fMRI motion estimates can be used as a surrogate score in the absence of better measures to control for motion in statistical analyses.

1. Introduction

Magnetic resonance imaging (MRI) of the brain is recognized as the gold-standard for in-vivo analysis of the central nervous system, its organisation, function, and diagnosis. The advent of high-quality structural 3D MR acquisition and sensitive software tools for its morphometric analysis have enabled accurate quantification of subtle neuroanatomical changes (Fischl, 2012; Henschel et al., 2022; Jenkinson et al., 2012; Reuter et al., 2012), e.g. in large cohort studies and clinical trials. While these neuromorphometric association studies commonly control for confounding variables and demographics, they frequently omit potential confounders of the MR acquisition (such as sequence, hardware, or firmware differences). This is especially problematic if confounding variables correlate with variables of interest and may bias statistical analysis. One such confounder is head motion. Head motion has been shown to systematically bias MRI-based neu-

romorphometric analysis (Alexander-Bloch et al., 2016; Biller et al., 2015; Gilmore et al., 2021; Nakamura et al., 2014; Pardoe et al., 2016; Reuter et al., 2015; Rosen et al., 2018; Savalia et al., 2017; Streitbürger et al., 2012). Specifically, motion-induced imaging artefacts can result in spurious findings of reduced gray matter volume estimates in T1-weighted MRI analysis (Alexander-Bloch et al., 2016; Gilmore et al., 2021; Pardoe et al., 2016; Reuter et al., 2015; Rosen et al., 2018; Savalia et al., 2017), even when they are barely visually detectable. Such motion induced biases are expected to increase with the introduction of higher resolution imaging sequences (Stucht et al., 2015) and extended scan times per sequence. Moreover, increased head motion is often correlated with variables of interest such as age (Madan, 2018; Savalia et al., 2017), body mass index (BMI) (Beyer et al. (2017, 2020), and various neurological disorders including autism (Cox et al., 2017; Dosenbach et al., 2017; Torres and Denisova, 2016), Parkinson's (Thenganatt and Jankovic, 2016), Alzheimer's (Versluis et al.,

* Corresponding author.

E-mail address: martin.reuter@dzne.de (M. Reuter).

<https://doi.org/10.1016/j.neuroimage.2023.120176>.

Received 28 February 2023; Received in revised form 22 April 2023; Accepted 15 May 2023

Available online 18 May 2023.

1053-8119/© 2023 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)

2010), and Huntington's disease (Rizk-Jackson et al., 2011), potentially resulting in an overestimation of these effects.

Common practice, such as the attempts to reduce motion by verbal instructions and cushioning of participants during acquisition, are highly recommended, yet, cannot completely control head motion. Retrospective expert grading of motion, based on imaging artifacts, can help identify severe motion cases, leading to their exclusion, but will not address consistent biases induced by small head motion. In contrast, the explicit inclusion of motion measurements as a control variable in statistical models offers a way to control even small motion biases and can help to disentangle motion effects from real anatomical changes. This, however, requires an accurate measurement of motion during MR acquisition in the first place.

In order to address this demand, we propose a *robust registration method* and an aggregated *motion score* for highly accurate head tracking and motion quantification during MR acquisition via a commercially available depth camera. For this, our registration yields a time series of head poses (*motion trace*) by robustly registering individual point cloud data derived from depth video frames to a reference frame. The resulting motion trace is highly accurate and able to recover high-frequency details. Our method significantly outperforms the vendor-supplied head tracking method (Slipsager et al. 2019, 2022) on three novel evaluations: 1. increased similarity to functional MRI motion estimates, 2. larger mutual information with a high frequency respiratory measurements, and 3. improved correlation of our motion score with various image quality metrics.

Finally, we apply our method in the Rhineland Study (Breteler et al., 2014; Stöcker, 2016), which is the first time optical tracking is used in a large population cohort study. While other work focuses on datasets with high motion levels, our method tackles the head tracking task in a cohort from the general population, where motion levels are low (requiring high sensitivity) and tracking needs to work on hundreds of participants with minimal intervention. We use our method to confirm previously known correlations of BMI and age with motion in the Rhineland Study. Additionally, we establish, that head motion increases over the course of the scan session, and that motion measured from functional MRI (fMRI) acquisitions can serve as a reasonable proxy of motion in adjacent scans.

1.1. Head motion estimation

As previously motivated, neuroimaging studies require quantitative head motion estimates to statistically model motion-induced biases and correct for correlates of increased head motion. These motion estimates can either be obtained during acquisition or retrospectively estimated from acquired images (Pollak et al., 2023). In practice, head motion is a composite of different motion patterns ranging from abrupt, large movements to slow drift and periodical patterns such as breathing (Zaitsev et al., 2015). To capture all types of motion (including small and rapid motion), a head tracking with high accuracy and high frequency is required. Coughing and diseases related burst movements can occur within a fraction of a second (Dbouk and Drikakis, 2020) or milliseconds (Eberhardt and Topka, 2017) and might not be detectable in motion tracking with sampling rates at 2 Hz or below.

Multiple recording and processing paradigms have been established for head tracking. They either directly utilize the MRI scanner or introduce external devices and are subject to individual trade-offs with respect to ease of use, accuracy, and sampling frequency. The most widespread approach is to register and align subsequent images (frames) of fMRI sequences. While its primary purpose is to correct inter-frame motion to establish frame-to-frame correspondences (Goto et al., 2015; Jenkinson et al., 2002; Maknojia et al., 2019), individual frame-to-frame alignments yield low-frequency motion estimates for the time frame of the fMRI sequence. Depending on its specific acquisition parameters, individual frames are acquired over 0.5 to several seconds aggregating the head motion across the frame and limiting the sam-

pling rate and ability to detect short bursts of motion. Unfortunately, the acquisition and alignment of fMRI images itself is also affected by motion artifacts induced by intra-frame motion, resulting in unreliable estimates for rapid motion (Maknojia et al., 2019). Finally, the multi-frame nature of the approach makes it inapplicable to single-frame MRI modalities, such as T1-weighted (T1w) structural scans and others. Based on the assumption of similar motion between functional scans and adjacent acquisitions, the fMRI-based motion estimate is sometimes used as a proxy for motion in adjacent structural scans (Pardoe et al., 2016; Savalia et al., 2017). While it has been shown, that ranking participants fMRI-based motion estimates is stable over multiple fMRI sequences in the same session, a direct comparison with motion during structural imaging sequences has not been performed in previous work.

Alternative MR scanner-based approaches rely on dedicated navigator sequences or signals that can be divided into three categories “FID”, “k-space”, and “image-based” navigators (Andronesi et al., 2021). FID navigators (Kober et al., 2011) estimate motion from sensitivity profiles of multiple coils rapidly. However, the methods require calibration or modeling for quantitative tracking and have limited resolution in the acquired data (Andronesi et al., 2021). K-space navigators, like navigator echoes (Costa et al., 2005; Fu et al., 1995), are used to perform real-time and inter-scan motion correction (Andronesi et al., 2021), which resulted in multiple methods (Van der Kouwe et al., 2006; Welch et al., 2002) with varying trade-offs (Andronesi et al., 2021) between accuracy and acquisition time. Image-based navigators acquire small volume frames, such as echo planar sequences (EPI) (Alhamud et al., 2012; Hess et al., 2011; Tisdall et al., 2012) that are acquired interleaved with other modalities and can be used to estimate motion via image registration. This estimation method is limited by the low image resolution and frequency of frames, and can increase the overall scan time (Andronesi et al., 2021). None of the scanner-based methods are currently widely used and scanner manufacturers move toward supporting external optical tracking systems (Andronesi et al., 2021), due to their higher sampling rates and increased accuracy.

Approaches based on optical tracking commonly detect the motion of markers, which are either fixed to the skin (Forman et al., 2011; Maclaren et al., 2012) or integrated into a dental brace (Maclaren et al., 2012; Stucht et al., 2015; Todd et al., 2015). Since optical systems operate independently of the scanner, they can capture high-frequency estimates for all MRI modalities without requiring integration into the acquisition protocol – while providing higher frequency and lower latency estimates than navigators (Aksoy et al., 2017). However, marker attachment may result in increased motion (e.g. saliva build up due to dental brace (Maclaren et al., 2012)) and longer scan preparation, which limits their applicability in high-throughput settings (Herbst et al., 2013).

To mitigate this limitations, recent work introduced markerless tracking of portions of the face with external cameras (Kyme et al., 2020; Olesen et al., 2010; Pardoe et al., 2021; Slipsager et al., 2019). These methods use high quality cameras e.g. time of flight cameras that capture depth images, to infer the head pose. While markerless tracking represents an unobtrusive tracking paradigm, it is more susceptible to challenges posed by participant-specific features, e.g. facial hair, head size, proximity to camera, and eye or skin motion from facial expressions.

The only available software that enables motion tracking based on 3D face captures is the camera vendor software “TracSuite” (Bergström and Edlund, 2014). In addition to scanner-based and optical tracking methods, EEG (Laustsen et al., 2022) and sensing pads (Musa et al., 2022) have very recently been proposed for within-scanner high-frequency head motion tracking. These approaches are promising as they are not affected by facial expressions and do not require marker attachment. Their tracking accuracy, however, is not yet well established.

Generally, the lack of standardized evaluation metrics and experimental procedures limits the comparability of approaches across pub-

lications (Kyme et al., 2020; Laustsen et al., 2022; Maclaren et al., 2012; Musa et al., 2022; Pardoe et al., 2021; Slipsager et al., 2019; Todd et al., 2015). Especially prominent marker-based (Maclaren et al., 2012) and markerless (Slipsager et al., 2019) optical tracking methods are primarily validated with MRI motion correction capabilities, which makes comparison challenging. Besides different sequences and correction algorithms, the cohorts vary drastically from diseased pediatric patients (Slipsager et al., 2019) to healthy adults with extensive MRI experience (Maclaren et al., 2012).

On the other hand validations on phantoms with induced motion (Einspänner et al., 2022) are repeatable, and can be adjusted for varying motion sizes and patterns. These movable phantoms, however, are not widely available and currently unable to model all aspects relevant to motion tracking, like non-rigid facial expressions or the wide range of motion types (such as breathing, swallowing, coughing, drift). Human studies with induced/requested motion can be used to study strong head motion for method development. But these studies lack both repeatability and realistic motion levels – making across study comparisons extremely challenging. Unlike phantom studies, human studies can not be used for direct validation as the real motion (ground truth) is unknown. Due to these challenges, many head tracking approaches are not well validated, highlighting the need for multiple secondary metrics to test the validity of within-scanner head tracking approaches for human studies.

1.2. Head motion as a variable of interest

Beyond the confounding effect of motion on morphological analyses, head motion in the MRI scanner is subject to numerous research that aims to uncover the relationship between increased head movement and other biological markers. Motion has been shown to correlate with measures of impulsivity/hyperactivity, diabetes, hypertension, nicotine, alcohol use (Couvry-Duchesne et al., 2016; Ekhtiari et al., 2019; Hodgson et al., 2017; Kong et al., 2014; Wylie et al., 2014), and various neurological disorders (Cox et al., 2017; Rizk-Jackson et al., 2011; Thenganatt and Jankovic, 2016; Torres and Denisova, 2016). A prominent correlate of within-scan motion is age, where head motion is increased in younger pediatric cohorts (Afacan et al., 2016; Barkovich et al., 2019; Dosenbach et al., 2017; Janos et al., 2019; Oztek et al., 2020), but decreased in younger adult cohorts (Andrews-Hanna et al., 2007; Chan et al., 2014; Madan, 2018; Pardoe et al., 2016; Savalia et al., 2017). Another strong correlate of within-scanner motion is the body mass index (BMI) (Beyer et al. (2017, 2020); Couvry-Duchesne et al., 2014; Ekhtiari et al., 2019; Hodgson et al., 2017; Kong et al., 2014) with a recent longitudinal study even showing the causality between weight loss and reduced within-scan motion in obese adults (Beyer et al., 2020).

On top of these correlates of in-scanner head motion, multiple independent studies (Couvry-Duchesne et al., 2014; Engelhardt et al., 2017; Hodgson et al., 2017) established in-scanner head motion as a heritable trait. Zeng et al. (Zeng et al., 2014) suggest that the association of reduced long-range connectivity and head motion is partially explained by individual variability in functional organization, in addition to bias due to motion artifacts, which outlines the previously discussed entanglement of cause and effect of motion in MRI acquisitions.

1.3. Scoring image quality and head motion

To correct for motion in statistical models and to model the relationship of motion and disease, we desire a scalar motion score during MRI acquisitions. Previously, manual motion scores were obtained by eyeballing image artifacts, such as ringing or blurring, that may be caused by excessive motion during the scan. However, such visual inspection cannot reliably identify all motion artifacts affecting downstream analyses as motion effects can still be detected in high quality images that pass visual quality checks (Pollak et al., 2023; Reuter et al., 2015). The

Rhineland Study performs visual quality checks for all T1w images, with three possible ratings: (PASS, WARNING, FAIL). In the subset of images used in this work, none failed and only 2.2% were rated as warning, indicating a low level of visible artifacts. This outlines the fundamental difference between general- and clinical cohorts, where Slipsager et al. (Slipsager et al., 2020) report 2.0% of images to be non-diagnostic and 7.9% of images to be of decreased clinical interpretability. These numbers likely understate the difference in image quality, since raters in the Rhineland Study look for any type of small image artifact, that can be caused by subtle movement, while clinicians primarily focus on the brain areas and motion levels that are relevant to diagnosis.

To address the challenge of quantifying barely visible motion artifacts, a variety of metrics are derived as continuous scores from images or motion traces. *Image-based scores* (Callaghan et al., 2015; Esteban et al., 2017; Pannetier et al., 2016) derive measures of quality directly from the MR images, often without reference. While they capture motion artifacts in constrained settings of motion correction experiments (Callaghan et al., 2015; Pannetier et al., 2016), image based scores struggle to disentangle the causes of image quality and thus cannot reliably quantify motion, constraining their usefulness to correct the biasing effect of motion in statistical models. The challenge to directly estimate motion from an image is rooted in the complex interaction of head-motion and modern MRI acquisition protocols, which is, for example, affected by parallel imaging and the chosen k-space trajectories. Therefore, even metrics that are known to be correlated with motion for specific acquisition protocols and in specific cohorts (Zaca et al., 2018) are unlikely to generalize across all acquisitions, especially when motion levels may differ drastically between previously investigated clinical cohorts with tremor and general population cohorts. The advent of deep learning may offer an opportunity to capture the complex interactions between motion and acquisition, however, initial research is still limited to predicting the experts sentiment (Fantini et al., 2018); (Küstner et al. (2018a,b); Largent et al., 2021; Lei et al., 2022; Ma et al., 2020; Stpien et al., 2021; Sujit et al., 2019; Zukić et al., 2022), or perceptual quality metrics, based on simulated data (Sciarrà et al., 2022). In initial work we leverage the motion quantification pipeline presented here with continuous motion scores to alleviate the reliance on manual labels (Pollak et al., 2023).

Since *motion-based scores* are computed from motion traces, they inherently reflect the participant's motion unlike image-based scores. However, not all motion maps to image quality in the same way. Particularly, a line of work on prospective motion correction (Castella et al., 2018; Todd et al., 2015) proposes a metric, that calculates a weighted sum of head speed over planes of 2D MRI acquisition. The weights are determined by a numerical simulation of “impact” on image and show that for the investigated acquisition, motion at scan start or end is less detrimental to quality. Unfortunately the metric can not be generated for 3D MRI acquisitions and it is unclear how accurate the numerical simulation and definition of motion relate to image quality in real-world settings. In fact, the analysis of timing, i.e. when motion occurs, as well as motion patterns, e.g. breathing, abrupt motion or drift remains an open question.

1.4. Challenges

The field of within-scan motion tracking has yet to unify metrics and methods to associate head motion with human characteristics and image quality for retrospective analysis. Comparison of tracking method is challenging due to the absence of real-world ground truth motion measurements in most datasets.

We aim to establish an accurate and reliable within-scanner motion tracking method in the Rhineland Study (Breteler et al., 2014; Stöcker, 2016), an ongoing large population-based cohort study with mostly healthy participants. Multiple measures are taken to reduce head motion during the image acquisition, leading to overall low head-motion compared to clinical datasets, or datasets recorded with induced mo-

tion, where motion levels are markedly increased. For this purpose we rely on a high frequency markerless registration approach.

For reliable tracking of rapid and short motion events, multiple samples should be taken during the event. Since quick head adjustments or coughing can occur in a fraction of a second (Dbouk and Drikakis, 2020; Eberhardt and Topka, 2017), tracking frequency should be significantly faster. We recommend tracking speeds of more than 5 Hz. Slow methods with frequency of 2 Hz or less are unlikely to capture such events, since movement occurs either during recording of a single sampling point or between two recordings. Additionally, high sensitivity is required to quantify small movements, which are predominant in the mostly healthy cohort under investigation. Further, in a high-throughput environment, such as the Rhineland Study, we require a protocol with minimal setup to reduce the time participants spend in the MRI scanner. Ultimately, real-time performance is desirable to enable applications beyond retrospective data analysis, such as providing immediate feedback to operators during acquisition to repeat severely motion-affected scans, or for automatic online motion correction.

1.5. Contributions

In order to address these challenges, we propose a method for generating motion traces from depth images and aggregating them into a motion score. This pipeline consists of a dedicated robust registration, that generates motion traces and further post-processing steps, where traces are synchronized, smoothed, re-sampled and aggregated to yield a score of average motion. We outline this process in Fig. 2, where we show the building blocks of our processing pipeline as well as the method validation.

For the first time, we validate an optical tracking method on a large general population cohort with respect to three different indirect measures of head motion: fMRI motion traces (rigid transformations), respiration signal (scalar), and T1-weighted image quality measures. The introduced metrics can be applied to any within-scanner head tracking method and provide scalar scores to benchmark tracking performance.

Using these tests, we show that i) our method captures accurate motion traces, that are similar to motion traces from an established fMRI motion estimation toolbox, ii) our method captures high-frequency, respiratory signals as recorded on the participants chest, and iii) our motion score – from depth images during structural image acquisition – correlates with known structural image quality measures. We outperform the vendor-supplied registration method for markerless tracking on all three benchmarks. Furthermore, iv) we replicate previous findings of correlation between within-scanner head motion with age and body mass index, to demonstrate the applicability in the neuroimaging research domain. Finally, v) we show longitudinal, within-session effects of increasing motion with time (within sequences and overall scan time).

2. Materials and methods

2.1. Data acquisition

The data is acquired in the Rhineland Study (Breteler et al., 2014; Stöcker, 2016), a large population cohort study. The study invites participants from the area of Bonn, Germany, of age 30 and over and it is carried out in accordance with the recommendations of the International Council for Harmonisation's "Good Clinical Practice" standards (ICH-GCP). Written informed consent was obtained from all participants according to the Declaration of Helsinki. Approval was granted by the Ethics Committee of University Bonn.

During the Rhineland Study's 1-hour MRI session, the scanner acquires up to 7 different MRI sequences, including T1-weighted (T1w) images and fMRI. Simultaneous optical head motion- and respiration tracking is performed throughout the session.

The participants are scanned with a 3T Siemens MAGNETOM Prisma MRI scanner (Siemens Healthcare, Erlangen, Germany) equipped with a 64-channel head-neck coil. The T1w images have a 0.8 mm isotropic voxel size and are acquired using a multi-echo magnetization-prepared rapid gradient-echo (ME-MPRAGE) sequence (van der Kouwe et al., 2008) with 2D acceleration (Brenner et al., 2014) and elliptical sampling (Mugler III, 2014) (repetition time (TR) = 2560 ms, inversion time (TI) = 1100ms, flip angle = 7°, field of view (FOV) = 256 × 256 mm, voxel size = 0.8 mm isotropic). The rapid whole-brain resting-state fMRI acquisition (Stirnberg et al., 2017) consists of 1070 frames with a 2.4mm isotropic resolution, each captured in 0.53 ms (TR = 530 ms, TE = 30 ms, flip angle = 16°, FOV = 192 × 192 × 144 mm, voxel size = 2.4 mm isotropic).

In one of the Rhineland Study's scan centers, an infrared time-of-flight camera (TracInnovations, Copenhagen, Denmark) (Olesen et al., 2010, 2012; Slipsager et al., 2019), mounted on the head-coil of the scanner, captures depth images of a portion of the participant's face during the scan. The depth images (also represented as point clouds) give a 3D representation of part of the participants face, captured through the opening between head-coil and a mirror in front of the participant's eyes. While the camera position is adjustable on a rail in the superior-inferior direction for a participant-specific field of view, neither distance to the participant nor the left-right position can be changed. Therefore, captured facial regions vary depending on head-size and position.

Following the camera adjustment, the MRI operator defines a mask around the eye of the participant by drawing a polygon on a 2D grayscale image also captured by the motion camera. Additionally, TracSuite (the camera vendor software) (Bergström and Edlund, 2014) assists the operator in excluding areas unsuitable for the measurement of head motion (e.g. the head-coil). Fig. 1b) illustrates the cleaned reference point cloud resulting from this process. Throughout the session, the camera acquires depth-images at non-equidistant time points (approx. 32 frames per second), which are converted to point clouds using the camera's field of view. Due to storage limitations, we only save every fourth point cloud, resulting in a sampling rate of approx. 8 Hz. The vendor software TracSuite has functionality for real-time head tracking. In addition to the raw-point clouds we also save the head motion estimates of TracSuite for method comparison, as well as time-stamps and metadata.

The Rhineland Study also measures the respiration of participants during the whole MRI session with a respiration sensor (PERU 098, Siemens Healthineers, Erlangen), that rests on the participants chest and captures its movement.

To reduce the head motion during the scan session, MRI operators ensure that:

1. The head is stabilized by inflatable (preferred), or foam cushions.
2. Each session is interrupted by two short, scheduled breaks during which the operator speaks with the participant in the scanner.
3. Nature scenes are shown during the whole session (Greene et al., 2018; Huijbers et al., 2017; Madan, 2018; Vanderwal et al., 2015), except for the initial resting state functional MRI scan, when participants are asked to fixate a cross on the screen.
4. Participants are instructed to remain still and stay awake.

The fraction of sessions with simultaneously acquired motion data (at the time of preparation of this work) consists of a total of 573 participants with a male/female distribution (self-reported sex) of 266/207 and age-range from 30 to 95 years. For developing and optimizing registration variants, we randomly select 10 participants, whose data we exclude from the method comparisons. For downstream analysis, we exclude 79 participants that did not complete the whole protocol or where motion data is not available for all scans (yielding 494 cases).

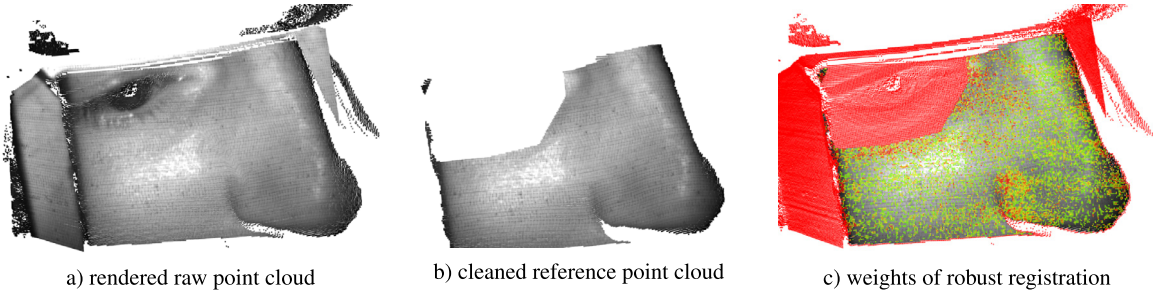


Fig. 1. Our method registers frames of the depth-image video (a) to the reference depth-image captured prior to acquisition (b). Our robust registration employs point-wise registration weights (c) to ensure less reliable depth-values impact the registration less (yellow) or not at all (red). These unreliable regions are predominantly located around the eye and on the head-coil, but also in areas of high noise and non-rigid motion.

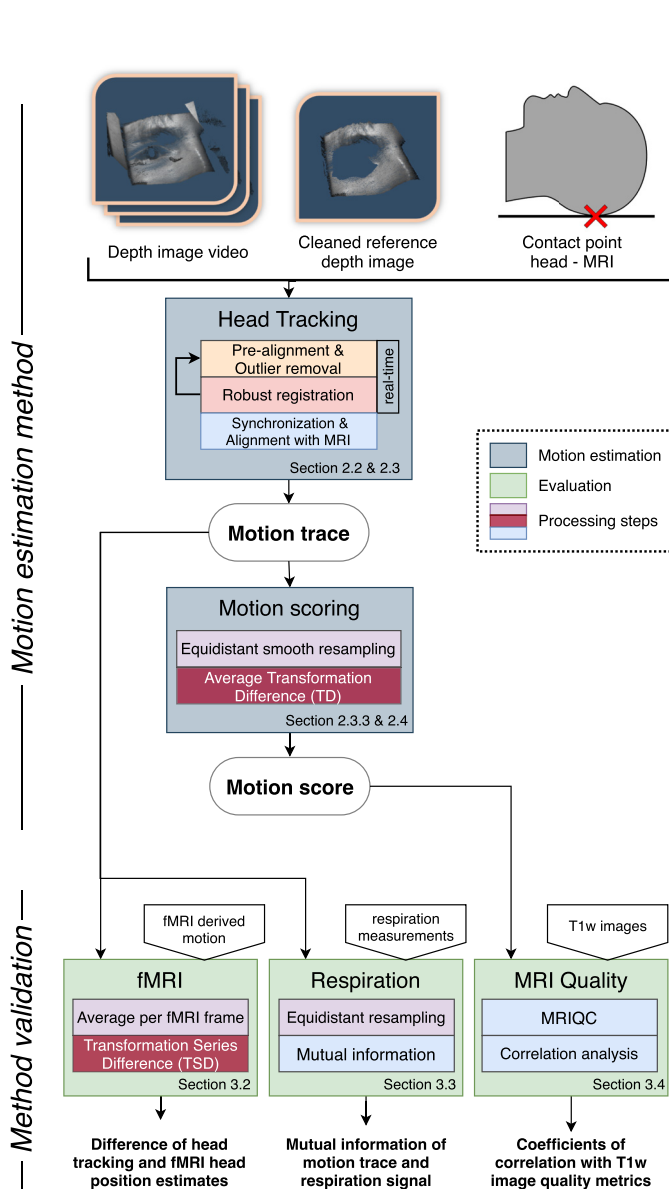


Fig. 2. Overview of the proposed method and pipeline to derive motion scores from depth images of the face. The two outputs of our method are i) a motion trace (i.e. series of head positions), describing head motion during the scan session and ii) a per-sequence summary motion score.

2.2. Motion estimation method

We compute a motion trace, by rigidly registering all point clouds (generated from depth images) to the manually cleaned reference point cloud. The result of the registration is the mapping of all captured point clouds to the reference, giving information how much the participants head moved during the scan. A classic solution for point-to-point registration problems is the Iterative-Closest-Point (ICP) algorithm (Arun et al., 1987). To address accuracy and robustness concerns of face registration such as noise and non-rigid facial skin motion, we introduce a robust criterion (following (Bergstr  m and Edlund, 2014)) as well as a regularizer.

2.2.1. Robust registration

We formulate the head tracking task as a sequential, rigid registration of point clouds captured during the scan session $P_k \in P_{1...N}$ to the reference point cloud P_{ref} . The resulting rigid body transformations consist of a rotation matrix $R_k \in \mathbb{R}^{3 \times 3}$, with $R_k^T R_k = I$ and $\det(R_k) = 1$, and a translation vector $t_k \in \mathbb{R}^3$. We call a time series of these rigid transformations (rotation and translation) a motion trace. The registration of $P_k \rightarrow P_{ref}$ can be characterized by solving the optimization problem:

$$R_k, t_k = \arg \min_{R_k, t_k} \sum_{p_k \in P_k} \psi \left(\underbrace{\min_{p_{ref} \in P_{ref}} \| R_k p_k + t_k - p_{ref} \|_2}_{:= res_p} \right) + \lambda. \quad (1)$$

In contrast to the unmodified ICP algorithm, where $\psi(res_p)$ is the identity function, we compare different robust criterion functions and, furthermore, introduce a regularizer $\lambda = \lambda(R_k, t_k)$ to discourage implausible solutions.

2.2.2. Robust criterion

Numerous candidates for the robust criterion $\psi(res_p)$ exist, such as Tukey Bi-weight, Cauchy-, Welsch-, and Huber functions (Bergstr  m and Edlund, 2014), which all down-weight or disregard point correspondences that do not satisfy the rigidity assumption based on the distance res_p to the nearest neighbour in the reference. We compare the criterion functions in Section 3.1 below and choose Huber's function as the best performer:

$$\psi(res_p) = \begin{cases} \frac{1}{2} res_p^2 & \text{if } res_p \leq Z \\ Z res_p - \frac{Z^2}{2} & \text{if } res_p > Z \end{cases} \quad (2)$$

$$Z = (1 + r_b) \text{median}(res_{1...N}),$$

where the hyper-parameter r_b determines a relative threshold beyond which the impact of large residuals is reduced and res_p is the residual associated with point $p \in P$. This robust approach also offers a solution to the partial registration problem, where a part of the registration object may be visible in one point cloud, but not in the other. While outliers in P_{ref} would never be considered due to the correspondence

search $\min_{p_{ref} \in P_{ref}}$, the robust criterion reduces the effect of outliers in $P_{1...N}$ by assigning them weights close to zero.

2.2.3. Regularization

In addition to the robustness criterion, we aim to incorporate known mechanics about within-scanner head motion into the registration. Because the camera only records a part of the face visible through the head-coil, the whole head pose is inferred based on features in the limited field of view of the camera at the front of the head (as seen in Fig. 1). This can reduce registration accuracy at the back of the head, where the skull is farthest from the visible region. Given that the head is resting on the pillow and tight padding prevents it from sliding, the contact point of the pillow and the back of the head can be encouraged to act as an additional (soft) anchor. Visual inspection of fMRI sequences confirmed that the back of the head is more stable than other regions, which is consistent with previous findings (Frost et al., 2019), and motivates this additional regularization.

The regularizing point p_{ref} at the back of the head is added to the optimization as a per-participant constant. We weigh the anchor with a weight of $\alpha = 0.03$ (a fraction fixed across all participants determined by experiments in Section 3.1) and multiply by the sum of all weights $\psi(\text{res}_p)$, resulting in the term:

$$\lambda(R_k, t_k) = \alpha \left(\sum_{p_k \in P_k} \psi(\text{res}_{p_k}) \right) (R_k p_{reg} + t_k - p_{reg}) \quad (3)$$

with additional variables in analogy to Eq. 1. We determine p_{ref} by first aligning the T1w image with a template (with FreeSurfer's robust register tool (Reuter et al., 2010)) and by then searching for the first voxels in the anterior direction. To get a reliable estimate, we ignore background noise by robustly conforming intensities and find the voxel closest to the median center of the cluster to use as the point p_{ref} . Finally, the regularizing point is mapped into the coordinate system of the motion camera as described in Section 2.3.1.

2.2.4. Performance and robustness considerations

To enable fast and robust sequential registration, we add two pre-processing steps, for each point cloud. These steps leverage the temporal consistency of the sequential point cloud data.

First, when registering the point cloud of the current frame to the reference, we initialize the rigid transformation with the result of the previous registration. This provides a good starting point for the registration optimizer and we only need to fine-tune the registration for the small movement possible within the sampling rate of the camera (approximately 125 ms in our case). Thus, the optimizer can converge with relatively few iterations, resulting in quick computation and additional robustness against unrealistically large movements (e.g. caused by measurement errors in a single frame).

Second, we discard distant outlier points prior to the optimizer loop. This is helpful, because shifts of head position can decrease the size of the overlapping region between the current and the reference point cloud. Such distant outlier points do not aid the registration, but affect the robust criterion through the median residual (see Eq. 2). We address this, by discarding distant outliers farther than 5mm from the reference point cloud, which enables us to keep the robust bound constant throughout all registrations. The registration accuracy is not affected, as the 5 mm distance threshold is conservatively estimated (about 30 standard deviations of mean motion in 125 ms). In addition, the a-priori removal speeds-up the overall computation.

Finally, to further increase registration speed, we under-sample the raw point clouds with a factor of three. This undersampling reduces head pose estimation time to approximately 110 ms, which is faster than the sampling interval of 125 ms per point cloud.

2.3. Post-processing

The time points inherited from point cloud samples, cause our motion traces to be non-equidistant in time and not always accurately

synchronized with the MRI scanner's internal clock. To facilitate the use of motion estimates in downstream analyses, we establish a post-processing pipeline that synchronizes all acquired MR images with the output transformations, reduces noise, and re-samples the transformations.

2.3.1. Aligning motion-tracker and MRI space

In order to compare head motion traces (e.g. camera-based traces with fMRI traces), they need to be defined with respect to a common coordinate system. Here, we choose the coordinate system of the participant's T1-weighted image. Since we can easily approximate the head center (and size) from the T1w image, the aforementioned mapping also enables us to accurately quantify motion within the head volume and to find the back of the head for regularization. To get the mapping between both coordinate spaces, we register the surface through the first high-intensity voxels of the T1w-image in posterior direction with the reference point cloud captured by the motion camera. To ensure good initial alignment, we pre-align the T1w images to a standardized template with FreeSurfer's robust register tool (Reuter et al., 2010).

The established correspondence between reference point cloud and the participant-specific anatomical space (T1w image) ties our motion tracking pipeline into common neuroimaging toolboxes, like FreeSurfer (Fischl, 2012). For example, registering functional MRI images to the T1w image with FreeSurfer's bbrregister (Greve and Fischl, 2009) enables a comparison of fMRI-based motion traces and motion traces from our method in a common space.

2.3.2. Time synchronization

The motion camera software (Slipsager et al., 2019) annotates individual point clouds with timestamps relative to the MRI session by requesting the time from scanner ad-hoc and correcting for latency. To address imperfect clock synchronization, we search for the offset (max. 15 seconds) that yields minimal differences between our and the fMRI motion traces (see Section 2.4.3). After synchronizing scanner and motion-tracking timestamps, our post-processing pipeline automatically reads meta-data from the scanner's DICOM files and annotates point clouds accordingly with the corresponding sequence and participant codes.

2.3.3. Re-sampling transformations

The time-stamps associated with the head poses determined by our registration are not equidistant, since the tracking system provides measurements at irregular intervals. Because post-processing steps, like motion scoring, require equidistant measurements at varying frequencies, we implement a re-sampling directly on the translation vectors and quaternions (Gramkow, 2001) of the motion trace by averaging over a sliding window with a triangular weight function. To parameterize this step for all applications we specify the window size and triangle slope separately, with a slope of zero resulting in an average along the window size (e.g. for averaging head poses during the acquisition of an fMRI frame).

2.4. Motion metrics

After determining the motion traces during a scan session and performing post-processing to prepare them for downstream analysis, we present multiple summary metrics to describe head motion in method validation and statistical analysis.

2.4.1. Head pose difference

For method validation and downstream analysis we require a metric that compares two head poses. While previous work compares head poses by translational and rotational errors of the head poses (Laustsen et al., 2022; Pardoe et al., 2021), we aim to combine head rotation and translation into a single score that weights both components according to the medical imaging application, as suggested by Wilke (2012, 2014).

Here, we employ the root mean square deviation for transformations (Jenkinson, 1999) (further *head pose difference*, HPD) to reduce the six degrees of freedom of a rigid transform (3 rotational and 3 translational) between two head poses into a single distance score. HPD is derived by averaging the displacements within a ball, which represents the head (Jenkinson, 1999). These displacements reflect the movement from one pose to another (described by transformation matrices). We approximate the parameters to this metric (head radius and location) from the population-average (distance from face to head center: 82.5 mm) and the individually placed scan origin, respectively. In practice, the latter is defined by the technician as part of the image acquisition for each participant.

The HPD distance measure is applied in two different contexts: 1. to quantify motion, i.e. to quantify how much the head pose changes between two measurements, and 2. to compare two transformations, e.g. for a comparison of methods.

2.4.2. Motion score

In contrast to pairwise pose differences, we also desire a robust summary motion score during a time interval (further *motion score* in millimeter per second), e.g., to quantify average motion for each MR sequence and include it into statistical analysis.

The proposed *motion score* is calculated by:

1. Re-sampling the motion trace (triangle slope = 0.1, window size = 9 samples), to create equidistant motion traces with head poses at exactly 8 Hz and to reduce measurement noise (see Section 2.3.3 Re-sampling).
2. Calculating the *head pose difference* between consecutive head poses and summing these differences for each one second interval (see previous section).
3. Averaging the one second *head pose differences* during the time of a single MR sequence.

The resulting summary scores are used as measure of average head volume displacement throughout each MR sequence.

2.4.3. Motion trace difference

For method validation, we require a metric that compares two motion traces (series of head poses). When one wants to compare two motion traces calculated from the same data (same sampling points), the *head pose difference* can be used on each pair directly. When comparing two motion traces acquired in the same time interval, but based on different samples and methods, we need to assert that

1. sampling time points agree,
2. transformations use the same coordinate space, and
3. comparisons are independent of the specific reference that was used in each method.

We ensure the first requirement of matching sampling points, by using the previously established re-sampling method. Second, we ensure a common coordinate space in the anatomical space of the structural MR scan as described in Section 2.3.1. The third challenge arises because head poses are always defined in comparison to a specific reference pose and these references can differ across methods (e.g. captured at different times, such as the beginning of the scan session or the beginning of the fMRI sequence). We, therefore, introduce a new metric *motion trace difference* (MTD) to compare motion traces from fMRI and optical tracking, that is independent of the specific reference pose. Thus, we compute transformations between each pair of poses for each method and average the HPD distance between corresponding transformations across methods (see Appendix D for details). This symmetric metric enables a fair comparison between two motion traces from different methods, with a summary measure of average displacement in the head.

In this work, we compare our motion traces to fMRI based motion traces, which are commonly used for in-MRI head motion analyses (Beyer et al., 2020; Dosenbach et al., 2017; Huijbers et al., 2017;

Pardoe et al., 2016; Savalia et al., 2017). While fMRI-based registrations are not without errors, we expect them to capture low-frequency head motion sufficiently well, especially since the whole skull is utilized for registration instead of only an area of the face as for the motion camera. We estimate the motion from fMRI with MCFLIRT (Jenkinson et al., 2002) and compare MCFLIRT's motion trace to our's by first averaging our motion trace over the length of the fMRI frame (570 ms). Then, we map both traces to the reference space and calculate the MTD as described above. While we consider all 1070 fMRI frames per participant in final results, 70 equally spaced fMRI frames sufficiently describe the performance characteristics for method development and testing in Section 3.1.

2.4.4. Mutual information of head motion and respiration

For additional sensitivity validation of head tracking methods, we aim to investigate how much of the densely sampled (256 Hz) respiration signal is contained in the head motion trace. The scalar respiration signal is measured by a respiration sensor attached to each participant's chest and indicates the respiratory phase and amplitude in the units of the sensor. As pre-processing, we re-sample the motion trace with our interpolation (see Section 2.3.3, triangle slope = 0.1 s⁻¹ and window size = 3 samples) and down-sample the respiration signal, with linear interpolation, to the point cloud measurement frequency of 8 Hz. Next, we compare the scalar respiration signal to motion traces with six degrees of freedom. We accomplish this by converting the transformation matrices of the motion trace into two vectors for translation and rotation (the latter being represented by quaternions). We ensure, that no sign flip occurs during conversion, by testing that the dot product of rotations in quaternion representation is positive and use all four components of the quaternion representation. The concatenation of both vectors results in a motion signal over time, that can be compared to a scalar respiration measurement. Finally, we estimate mutual information (Kraskov et al., 2004; Pedregosa et al., 2011) between the respiration signal and each of the motion signals. The sum of mutual information scores then represents how much information captured by the respiration sensor is contained in the motion trace.

2.5. Statistical methods

We require statistical tests for the validation of our methods as well as for the analysis of cross-sectional and longitudinal motion effects in a scan session. First, we perform a paired difference test across methods using the non-parametric Wilcoxon's signed-rank test. Second, to test whether two measures are linearly correlated, we use Pearson's-r (e.g. for comparison of motion with image quality metrics). For both tests we employ the implementations of the SciPy (Virtanen et al., 2020) Python library. To reproduce previous cross-sectional results, which show correlation of risk factors age and body mass index with motion, we use a least squares regression model with BMI, age, and sex as dependent variable and average motion over all scan sequences as the variable of interest.

Finally we test the hypothesis, that participant motion increases over the course of the scan session. For this task, we model motion using a longitudinal linear mixed effects model (LME). We include time as well as interaction terms of time with age, sex, and BMI. The model is described by the term (participant i , sampling point j at every 1 second interval):

$$Y_{i,j} = \beta_1 + \beta_2 t_{i,j} + \beta_3 \text{BMI}_i + \beta_4 \text{age}_i + \beta_5 \text{sex}_i + \beta_6 t_{i,j} \text{BMI}_i + \beta_7 t_{i,j} \text{age}_i + \beta_8 t_{i,j} \text{sex}_i + b_i + e_{i,j} \quad (4)$$

with b_i as subject specific random intercept, β_1 as global intercept and e_i the random error. The dependent variable $Y_{i,j}$ is motion (i.e. head speed) in (mm/s). t_i is the time from start of first image acquisition, BMI is the body mass index, and age the participant's age in years. The latter two markers are centered around zero prior to model fit. We also

include sex as a categorical variable. Data points are only sampled during scan sequences and not during breaks, where larger motion can occur. Both statistical models are implemented in the statsmodels Python library (Seabold and Perktold, 2010).

3. Results

In this section we compare our method to the state-of-the-art registration – the software of the motion cameras manufacturer (TracSuite (Slipsager et al., 2019, 2022), Version 3.0.0). We determine both methods' quality by measuring similarity to functional MRI motion and to respiratory traces. Further, we explore the correlation of motion scores with structural image quality measures as well as with known correlates of motion (BMI and age). These downstream correlation analyses serve as a validation experiment for the pipeline consisting of the registration method and the proposed motion score for statistical analysis. Finally, we apply our method in the Rhineland Study to investigate, whether participants move more in later MR scans during a scan protocol and whether fMRI motion estimates can be used as an approximation of motion for later sequences.

3.1. Ablation study

Prior to the comparison with another method, we perform an ablation study. This type of study investigates performance of a system by removing, adding, or exchanging individual components one-at-a-time to understand the contribution of each component to the overall system. Here, we investigate which choices are most important for the final method performance by running different variations of the robust registration and thus determine our final method for the subsequent analyses. To prevent leaking information about the test data into method development we use a separate set of 10 scan sessions for the optimization of our method and exclude these cases from subsequent evaluation.

The MTDs to MCFLIRT's motion traces (i.e. the difference to fMRI-based motion estimates) of the registration variations are shown in Fig. 3). We investigate the impact of the robust bound r_b , as well as the criterion function – both of which define the robustness of the iterative closest point registration. We find the best criterion function is Huber's function with $r_b = 0.2$. We test three other criterion functions, but for the same, optimized robust bound and same asymptotic variance of the functions, the difference in performance is small. A noticeable performance increase, however, occurs for the regularizer, whose introduction reduces the transformation series difference to the fMRI scan by more than 0.5 mm.

To show the trade-off between method speed and performance we vary the number of internal iterations of the robust iterative closest point algorithm. We find that there is no performance gain for more than 15 performed iterations, indicating fast convergence for most registrations. We conservatively choose 30 optimizer iterations for the final method. We also observe little performance loss for an under-sampling of point clouds on a grid up to every fourth point. This specific comparison, however, may be insensitive to random noise, due to the averaging of multiple motion camera based head poses for the comparison with the low-frequency fMRI based motion trace. In an additional evaluation with the respiration signal on the ablation data set (no dedicated Figure), we find that under-sampling decreases performance when looking at mutual information with respiratory measurements (1: 0.79, 2: 0.77, 3: 0.76, 4: 0.71; [undersampling factor: mutual information]) showcasing the importance of different secondary metrics for method validation. Since MI drops substantially for an under-sampling factor greater than three, we choose three as factor for the final method – which still allows for online registration.

Finally, we investigate how the reference point cloud affects the registration performance. We compare different pre-processing steps of the reference point cloud and find that omitting pre-processing (*raw pc*) reduces registration performance. Both smoothing and removing artifacts

as a pre-processing (*cut pc*) as well as the additional manual masking step (*cut pc no eye*) increase accuracy. As an additional experiment, we extract a reference point cloud from the T1w image, which contains the whole face (*T1w pc no eye*). In our case, this MR-based reference performs worse than cleaned references based on the depth camera data. The full view of the face may, however, prove to be an advantage for large motion cases, where overlap of visible face region with the camera reference could become too small.

3.2. Comparison to fMRI motion estimates

As a first method validation, we compare the multiple kinds of motion traces from the depth camera to motion traces based on resting state fMRI. The evaluation follows the same setup as our ablation study in the previous section, but is performed on the large dataset of 563 fMRI acquisitions.

The three evaluated methods are: 1. The trivial “no registration” method, which only predicts a series of identity matrices, 2. TracSuite – the camera manufacturers method (Slipsager et al., 2019), and 3. our registration method. We show the results of the analysis in Fig. 4. Our method is significantly closer to the fMRI motion estimates than TracSuite or assuming no motion ($p < .001$). This shows that our method captures similar movements as MCFLIRT from the independent fMRI acquisition. Surprisingly, TracSuite's estimates are significantly farther from fMRI based motion traces for the investigated population cohort than assuming no motion ($p < .001$).

3.3. Comparison to respiratory chest movements

We estimate the mutual information of motion traces and measurements of a respiration sensor on the participants chest. This comparison showcases the ability to capture subtle head movements, that can also be measured by a high-frequency respiration sensor. We test the mutual information between respiratory chest movements and 1. a random motion baseline (6 independent uniformly distributed random variables between 0 and 1), 2. TracSuite, and 3. our registration.

The mutual information between respiration- and motion measurements is shown in Fig. 5. Both head tracking methods capture information about the participant's breathing and chest movements. Our method significantly outperforms TracSuite in this validation ($p < .001$). For additional qualitative comparison, we show the motion traces and respiration signal of one representative sample in Appendix B.

3.4. Motion correlates with image quality metrics

To jointly validate our registration as well as our motion scoring method, we test the correlation of *motion scores* for the T1w image acquisition with different measures of structural image quality.

The image quality metrics are determined with the MRIQC (Esteban et al., 2017) toolbox and have been either associated with motion (Atkinson et al., 1997; Provins et al., 2023; Shehzad et al., 2015) or with perceived reduction in quality (Mortamet et al., 2009; Shehzad et al., 2015). Since the competing method, TracSuite, does not provide a scoring method, we apply our scoring method to its motion traces. The results are presented in Fig. 6, where the pair-wise linear correlation coefficient (Pearson's-r) is shown for five quality metrics, that are associated with motion and motion artifacts in structural images. The motion score based on our registration method has a higher correlation than the score based on TracSuite across all metrics.

3.5. Motion correlates with BMI and age

As a final validation, we test whether we can find known correlates of head motion with the combination of our registration method and motion scoring. Increased head motion is known to be associated with

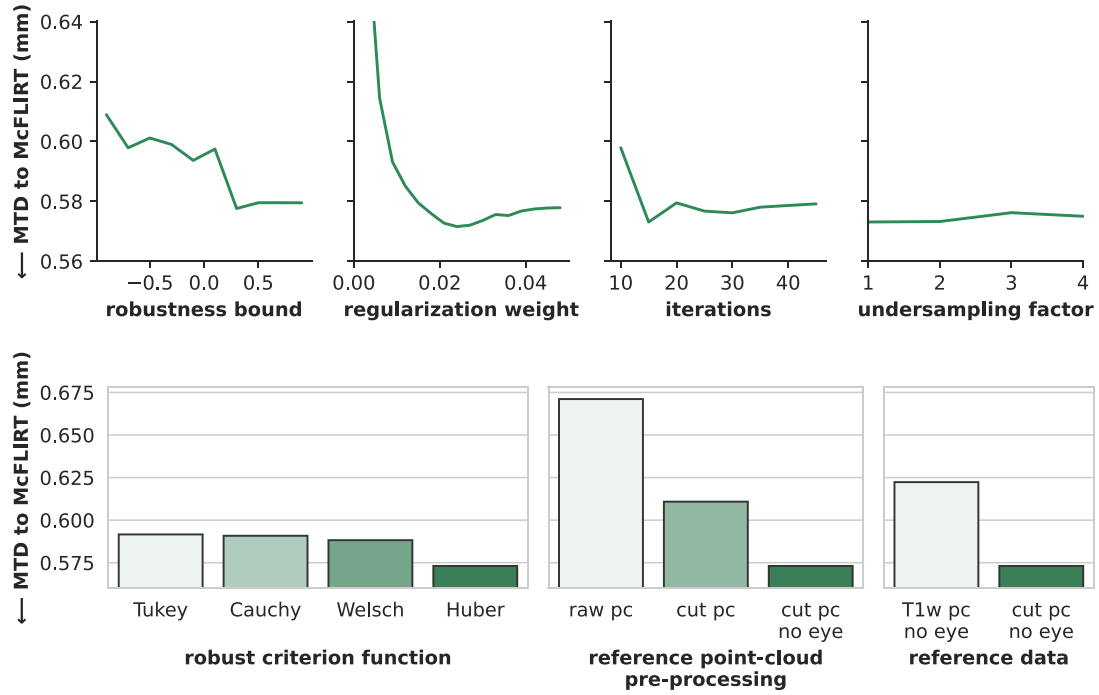


Fig. 3. We test different configurations of our registration method in an ablation, using the transformation series difference (MTD, see Section 2.4.3, lower is better) to a fMRI based motion traces of McFLIRT as metric. The top row depicts results for continuous parameters of the registration (left to right): the robust bound r_b , relative weight of regularization, the number of iterations of the robust registration, and the under-sampling factor of the input point clouds on a grid. The bottom row shows four different robustness criterion functions (Bergström and Edlund, 2014) (left), and four different reference point clouds (right) including a raw point cloud (*raw pc*), a point cloud generated from the T1w image with the eye cropped out (*T1w pc no eye*), the cleaned point cloud with cropped out head-coil and smoothing (*cut pc*), and the cleaned point cloud where the eye is additionally cropped out (*cut pc no eye*).

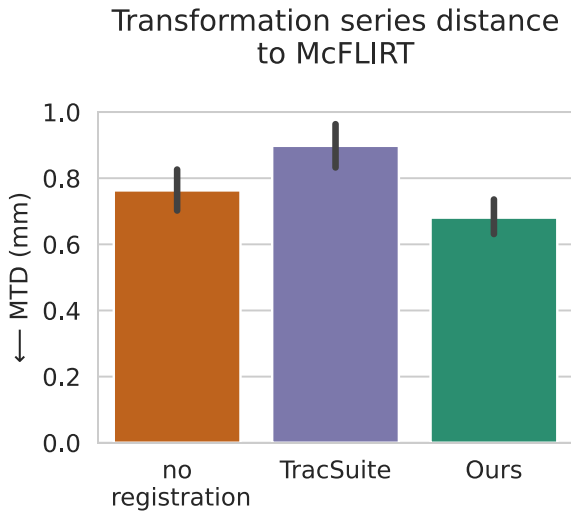


Fig. 4. Comparison of registration methods with fMRI motion traces as reference standard. The y-axis shows the transformation series difference (MTD) to motion traces of McFLIRT. “no registration” describes a motion trace, where zero motion is estimated (identity matrices). All pairwise differences between methods are significant, measured by the Wilcoxon signed-rank test ($p < .001$).

increased body mass index (BMI) (Beyer et al., 2017, 2020; Couvy-Duchesne et al., 2014; Ekhtiari et al., 2019; Hodgson et al., 2017; Kong et al., 2014) and with age (Andrews-Hanna et al., 2007; Chan et al., 2014; Madan, 2018; Pardoe et al., 2016; Savalia et al., 2017). To test whether we can replicate these cross-sectional effects with motion measured by our method, we fit a linear regression model with average motion across all scans in the session as the dependent variable and the independent variables BMI, age, and sex. The correlation of both age ($r = 0.0023, p < .001$) and BMI ($r = 0.0053, p < .001$) with participant

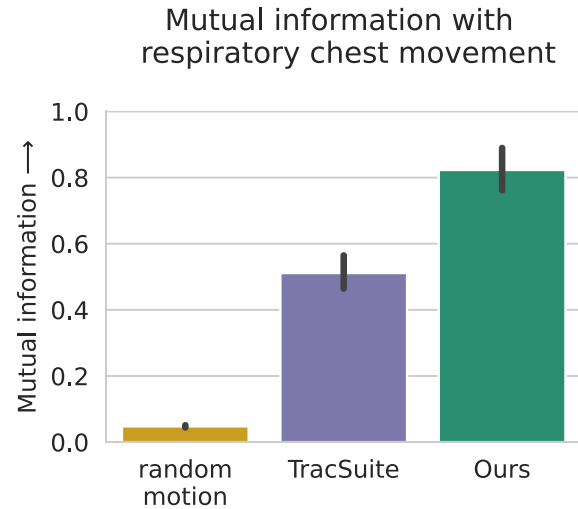


Fig. 5. Comparison of registration methods by mutual information (MI) with respiratory chest sensor movements. High MI indicates that chest movements are contained in the head motion traces. Random motion describes a randomly generated transformation series. All pairwise differences between methods are significant, measured by the Wilcoxon signed-rank test ($p < .001$).

motion is highly significant (see Fig. 8), confirming previous findings with our method. Participants sex is not significantly correlated with in-scanner head motion ($r = 0.0089, p = .446$). These expected findings showcase, that our metric can be used to quantify motion for downstream analysis.

3.6. Longitudinal motion effects during a scan session

Finally, we apply our method by modeling motion effects and interactions over the course of the whole scan session. Since the optical

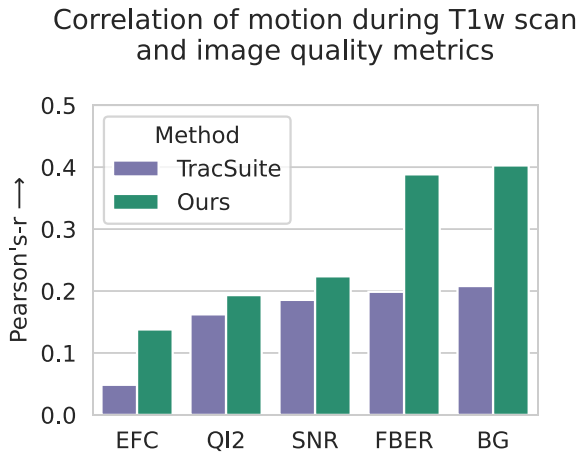


Fig. 6. Comparison of registration methods by linear correlation coefficient (Pearson's-r) of calculated motion score and T1w image quality metrics. Since competing methods do not generate a motion average, we use our motion score to aggregate both motion traces. The shown metrics generated by the MRIQC (Esteban et al., 2017) toolbox are: the entropy focus criterion (EFC) (Atkinson et al., 1997), Mortamet's quality index 2 (QI2) (Mortamet et al., 2009), signal to noise ratio (SNR), foreground-to-background energy ratio (FBER) (Shehzad et al., 2015), and mean background intensity (BG).

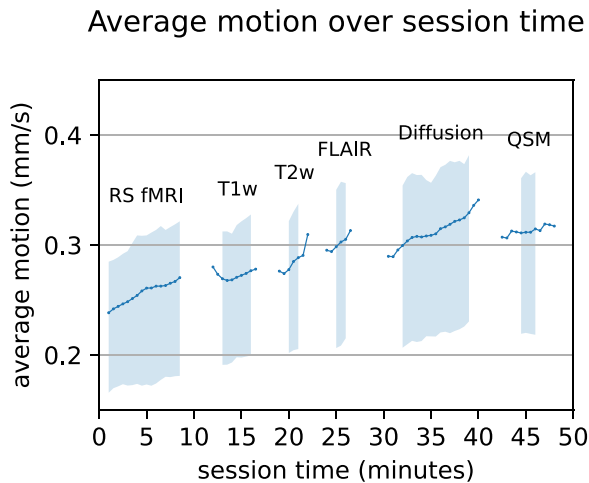


Fig. 7. Motion (30-seconds moving average) increases during image acquisition and throughout the full MRI session. The line indicates the average of participants motion, if samples of more than 20% of participants are available and the shaded area indicates the interquartile range if samples of more than 80% of participants are available. The number of available samples differs over time due to varying gaps between sequences. Annotations indicate the different MR sequences.

tracking method is independent of the scanner we can, for the first time, directly compare motion in a 1h scan session across multiple MR sequences. We perform this analysis with the linear mixed effects model (LME) described in Section 2.5.

In the standardized protocol of the Rhineland Study, motion increases significantly over time with approximately 0.6% per minute (0.001751 mm/s) of mean motion across all participants [MMAAP] (time: $p < .001$, $z = 706.145$) as illustrated in Fig. 7. A similar increase can be observed per one year of participant age (0.7% MMAAP, 0.002077 mm/s, $p < .001$, $z = 5.159$). We also find a strong correlation with the body mass index (BMI), where one point of BMI increased motion levels by around 1.5% MMAAP (0.004397 mm/s, $p < .001$, $z =$

3.411). There is no significant difference in motion levels across sexes ($z = 0.493$, $p = .622$).

The linear mixed effects model also contains time interactions. We observe a significantly higher increase of motion over time for increased age and body mass index of participants (time $\times$ BMI: 0.021% MMAAP 0.000060 mm/s, $z = 132.906$, $p < .001$, time $\times$ age: 0.002% MMAAP, 0.000002 $z = 40.322$, $p < .001$). This indicates that levels do not only increase in groups of higher BMI and age, but that the increase of motion over time in the scanner is amplified by these correlates of head motion, though the additional increase is very small in comparison to the general increase of motion during the session. Participant sex has a larger measured impact on motion increase over time with male participants having a slower increase of motion over time (time $\times$ sex: 0.069% MMAAP, 0.000199 mm/s, $z = 52.392$, $p < .001$), despite no significant cross-sectional effect. All of the described interaction effects, however, are weak compared to the overall increase over time, suggesting that cross-sectional motion differences are similar throughout the scan sessions.

Since within-scanner motion tracking is not always available, we test whether motion can be inferred from one sequence to later sequences. A previously proposed motion correction strategy is using readily available motion traces from fMRI to exclude participants with increased motion levels (Pardoe et al., 2016; Savalia et al., 2017) from morphological statistical analyses in order to reduce bias. In the Rhineland Study's standardized protocol the resting state acquisition is first in the session. For the first time, we compare direct estimates of motion during functional MRI and other MRI sequences, to see whether motion levels are stable throughout the scan session. In a paired linear correlation analysis, shown in Fig. 9, we observe a high correlation coefficient for all scans in the one hour protocol (Pearson's-r; adjacent to fMRI: $t = 0.88$, $p < .001$; 35 minutes after fMRI: $t = 0.77$, $p < .001$).

4. Discussion

We present a robust method to extract motion traces and a representative motion score from a depth video of the face. The proposed registration method outperforms the vendor-supplied method on three indirect validation tasks. Besides the registration method we introduce tools to compare head motion traces. We, furthermore, perform association studies of head motion as a variable of interest with BMI, sex, age and scan session time in the Rhineland study.

The proposed registration method combines a robust point-to-point registration with a-priori outlier detection and a regularizer specific to the face tracking application. A key to the performance and robustness of our method is exploiting the high sampling frequency of the camera. Contrary to e.g. fMRI based motion tracking, the movement performed between measurements is much smaller, allowing for a detailed estimation of the head motion. This also allows a highly constrained registration, which targets a smaller range of movements possible in 125 ms and permits real time processing. Markerless tracking, however, comes with a trade-off. Contrary to MRI-based motion tracking, only a small part of the head is visible and the skin can cause non-rigid movements. We address these limitations with two additions to the registration. To address non-rigid movements and noise on the measured face area we add a robustness criterion (Bergström and Edlund, 2014) and to achieve stable inference of the head position at the back of the head, we introduce an anatomically motivated regularizer. We evaluate the impact of these additions and their parameters in the ablations study (Fig. 3) and see large improvements for sensible robustness criteria and regularization.

Direct method validation with ground truth motion is exceedingly rare and difficult in the head motion tracking domain. A promising avenue for this are movable head phantoms (Einspänner et al., 2022). However, these phantoms are not widely available and can not yet imitate realistic movement patterns of participants. Therefore secondary metrics of head motion, like fMRI-based estimates should be chosen for indirect method validation. We believe that multiple secondary metrics

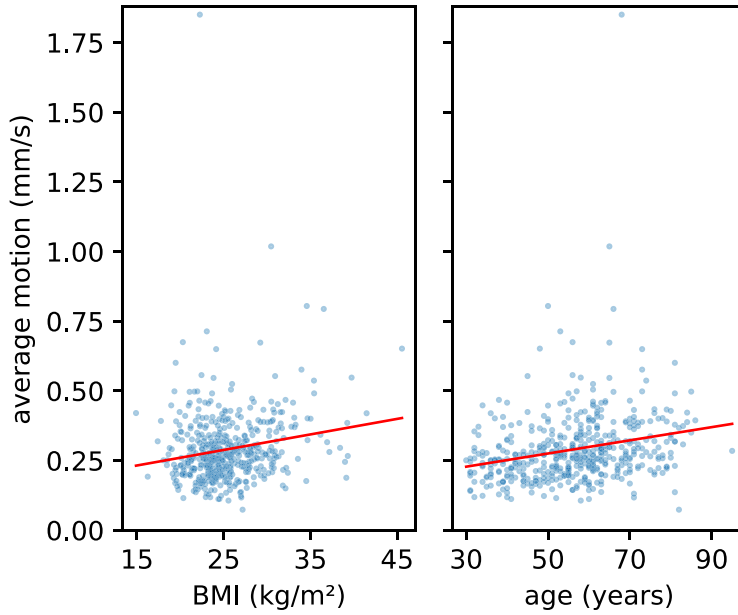


Fig. 8. Our average motion score correlates significantly with body mass index (BMI) and age in a linear correlation model, that includes sex, BMI, and age ($p < .001$). Depicted is the scatter data with independent linear fits.

Correlation of 2-minute average motion with avg. fMRI motion

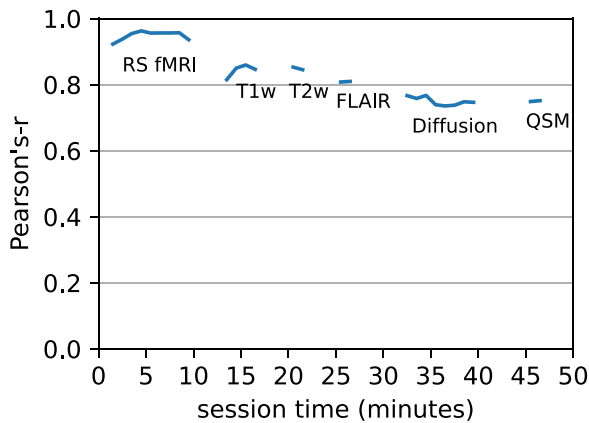


Fig. 9. Paired linear correlation (Pearson's-r) analysis: motion during a resting state fMRI (RS fMRI) sequence is a good approximation for depth camera motion estimates (1-minute moving averages) during following sequences. The measures were averaged in 1 minute bins and bins with less than 90% of total available samples were discarded. For a description of sequences and correlations per image acquisition see A.1.

are required for method validation in order to reduce bias in the comparison of two imperfect motion estimates. Unfortunately, this practice is currently not widespread, with some previous markerless and marker-based optical tracking methods only using MRI motion correction capabilities as real-world validation (Maclaren et al., 2012; Slipsager et al., 2019). With the proposed extension of Jenkinson's transformation difference to motion traces and the use of three secondary sources of motion measurements: 1. fMRI-based motion traces, 2. respiratory chest movement, 3. structural image quality metrics, we hope to contribute to the development and validation of head tracking methods in general.

A limitation of our test setup is the storage space for motion sequences, which only allows us to save every fourth point cloud for retrospective registration. The vendor-supplied software "TracSuite" processes all point clouds directly upon capture and, therefore, has access to additional point clouds. We expect this asymmetry of conditions in the method comparison with otherwise same conditions to favor Trac-

Suite, since more point clouds can be used to reduce the acquisition noise (which was a key challenge throughout the development of this method). Future applications of our method executed in parallel with the MR acquisition could also benefit from the additional temporal resolution. Furthermore, it would allow: i) using our metrics and accurate tracking to provide real-time feedback about participant motion to MRI operators, and ii) improving previously proposed (Frost et al., 2019; Slipsager et al., 2022) markerless online motion correction methods.

Our method outperforms TracSuite on all three validations. While TracSuite has high mutual information with respiratory chest movement, matching the findings of (Slipsager et al., 2019), it performs worse than no registration (identity) in comparison to fMRI based estimates for our 563 participant dataset. These inconsistent results underline the importance of multiple different metrics when performing indirect method validation. The surprising under-performance of TracSuite might be an indicator that it targets high-motion, clinical groups and does not generalize well to compliant participants with little head motion, where no motion is a decent assumption. Since the competing method is not open source, we cannot further investigate this result. The overall low motion in our dataset can be seen by the seemingly high performance of the "no registration baseline". Participants remain still for large parts of the scan causing the residual MTD to MCFLIRT of the baseline and our method to be of similar magnitude. At short times, the difference between the two methods, can be larger than the head motion. However, when averaging over the whole dataset, our method agrees with MCFLIRT's estimates of head position. To showcase the difference to high-motion datasets we show the comparison to fMRI for a case of extreme, induced motion in Appendix C, where both methods clearly outperform the baseline. A strength of our methods is that the fMRI validation can be performed without additional hardware and therefore our method can easily be re-optimized and validated for different scan setups and cohorts. For the extreme motion case, for example, we re-ran the optimization and found that regularization is not helpful in large motion settings, but exchanging the motion-tracker derived reference by the reference derived from the T1w MR-image improves the registration performance for this case, which has very different, larger patterns of movement, due to induced motion and removed padding.

While the estimation of head positions is optimized with fMRI-based estimates as ground truth, our score of average motion is theoretically motivated. In this work, we do not consider the way motion impacts MRI scans differently depending on the type and time-point of motion. Nonetheless, we can show correlation with image quality metrics

and achieve higher correlation coefficients than the competing method in a fair comparison (using the same scoring). This implies that our method quantifies motion relevant for acquisition more accurately than the other methods.

The strength of the proposed scoring is the direct application as scanner-independent metric for statistical analyses in population studies. We show this by performing a straightforward statistical analysis to replicate previous findings of motion associations with BMI and age and, furthermore, apply our method to test whether motion increases throughout the scan sessions. We find that time in the scanner is indeed a strong, significant correlate of motion, confirming previous fMRI-based studies of in-scanner head motion (Meissner et al., 2020). Motion is, for example, increased by approximately 29% after 35 minutes of scanning in our dataset. This trend can also be observed in Fig. 7, where we show motion increases over time. A confounder of this visualization is a reduced sample size at the start and end of MRI sequences. Since breaks between scans are inevitably of different length, the start and end of each sequence include fewer samples, containing participants with exceptionally short or long breaks. We outline this effect, by removing the confidence interval and plotted line for fewer than 20% and 80%, respectively. The resulting graph shows an intra-scan and intra-session increase of motion, but an inter-scan reduction of motion in most cases, which is in agreement with the previously demonstrated effectiveness of short breaks (Meissner et al., 2020). We also note, that motion measurements seem to increase more than expected during the Diffusion weighted image (DWI) acquisition and decreases afterwards. This phenomenon might be connected to the rapidly changing gradients during DWI, which cause loud noises and vibrations. This, in turn, could affect participants' head motion levels and possibly also the measurement system itself. We, however, did not find any noticeable vibration artifacts in the motion tracking point cloud data by visual inspection.

Despite a marked increase of motion over time, overall motion levels can be considered low. The Rhineland Study with its compliant and predominantly healthy participants incorporates several procedures to reduce head motion, which results in small absolute effect sizes. This is corroborated by high structural image quality (shown by the aforementioned low number of WARN or FAIL ratings compared to clinical cohorts) and the fact that “no motion” is a decent estimate for fMRI based head motion traces, while the difference is higher by an order of magnitude in an induced motion experiment (see Appendix C).

Nevertheless, the LME also reconfirms significant cross-sectional associations of BMI and age with motion. Given that motion can bias downstream morphometric analyses (Alexander-Bloch et al., 2016; Reuter et al., 2015), it is likely beneficial to control for head motion, even when it is small, to disentangle motion effects from real anatomical changes. The age correlation further underlines that quantifying and controlling for head motion could be important when aiming to find subtle longitudinal differences.

Age and BMI are only two of the many potential markers that correlate with motion. We did not yet investigate any disease states, smoking, sleep, sport, or medications, which is an obvious avenue for future work. Even beyond, high-frequency motion tracking can be used to analyze motion patterns of participant groups. Previous research, for example, has shown causation between increased BMI and increased motion levels (Beyer et al., 2020) with fMRI based motion estimates, but it remains unclear what type of movements were exhibited by overweight participants. The authors cite alterations in the respiratory system (Littleton, 2012) and alterations in dopaminergic signaling (Tomasi and Volkow, 2013) as possible reasons for a decrease of head motion after drastic weight loss, due to bariatric surgery. As we show in our validation, the proposed method captures head motion associated with respiration, in addition to larger, impulsive movements. By measuring the change in these different motion types, researchers could gain valuable information in the possible mechanisms of increased head motion.

We note, that sex is not significantly associated with motion in the investigated population, despite previously shown significant differences, with females moving less than males (Huijbers et al., 2017), especially for pediatric patients (Dosenbach et al., 2017; Frew et al., 2022; Madan, 2018; Pardoe et al., 2016). Surprisingly, we find that males' motion levels increase significantly less than females' motion levels during the scan session. We also find significant interactions between time in the scanner with BMI and age, with increased BMI and age associated with a faster increase of motion over time. These results indicate, that the potentially increasing discomfort is amplified in older participants and participants with larger BMI. The interaction effect sizes, however, are very small compared to the cross-sectional motion effects and the overall motion increase with time.

The small size of interaction effects indicates a high rank-consistency of motion levels during the scan session as previously shown during multiple functional MRI measurements (Savalia et al., 2017). Since motion tracking is not readily available on all sites, we investigate whether we can measure direct paired linear correlation between motion during the initial fMRI sequence and following sequences. High correlation indicates that a single measurement of motion at the beginning is predictive of motion levels, even at later stages of the scan protocol. Therefore an initial measurement could be enough to correct for motion in all sequences, given a consistent ordering of scans. Fig. 9 shows that correlation with average motion during resting state (RS) fMRI is highly correlated with two minute intervals during the same sequence, but the correlation drops sharply at the begin of the next sequence (T1w image acquisition) and then continuously decreases. This could be caused by a different scanning environment during functional scan (with a fixation cross, typical for RS fMRI) and other sequences where a movie is shown to participants. Watching movies has been shown to reduce head motion compared to a fixation task (Greene et al., 2018; Huijbers et al., 2017; Vanderwal et al., 2015), which might influence correlation of motion levels. Nonetheless, we see rather high paired linear correlation between average RS fMRI and the following sequences, indicating that, in absence of dedicated motion tracking, fMRI can be used to estimate motion levels during the whole scan session.

Due to the significant increase of motion over time, we recommend that important sequences or sequences with high sensitivity to motion should be acquired first - especially in long sessions. The common assumption of increased motion at the beginning of the first scan can be rejected in accordance with previous literature (Meissner et al., 2020). Further, to reduce motion bias, the order of sequence acquisition should ideally remain constant in longitudinal studies.

Overall, we introduced a performant, accurate method for robust registration of depth images of the face. This method, in combination with extensive validation and reliable aggregation into an average score, is capable of finding previously known and unknown motion effects in a large population cohort study. It, furthermore, provides a sensitive measure of head motion to analyze and control motion effects in statistical association studies.

Data & Code availability statement

This work uses data from the Rhineland Study. Data of the Rhineland Study is not publicly available because of data protection regulations. However, access can be provided to scientists in accordance with the Rhineland Study's Data Use and Access Policy. Requests to access the data should be directed to Dr. Monique Breteler at RS-DUAC@dzne.de. The source code for the registration and validation methods will be made public upon acceptance at <https://github.com/Deep-MI/head-motion-tools>.

Declaration of Competing Interest

The authors do not declare any conflicts of interest.

Credit authorship contribution statement

Clemens Pollak: Methodology, Data curation, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **David Kügler:** Conceptualization, Methodology, Investigation, Supervision, Project administration, Writing – original draft, Writing – review & editing. **Monique M.B. Breteler:** Funding acquisition, Resources, Writing – review & editing. **Martin Reuter:** Conceptualization, Methodology, Resources, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgments

We would like to thank the Rhineland Study group for supporting the data acquisition and management. We would like to thank especially Shahid Mohammad, Santiago Estrada and Valerie Lohner for their expertise around the Study's processes. We would also like to thank Kersten Diers for his consultation on statistical analyses and Rüdiger Stirnberg for supporting the acquisition of experimental data. This work was supported by DZNE institutional funds, by the Federal Ministry of Education and Research of Germany (031L0206, 01GQ1801), the Helmholtz-AI project DeGen (ZT-I-PF-5-078), and by NIH (R01 LM012719, R01 AG064027, R56 MH121426, and P41 EB030006).

Appendix A. Average motion levels

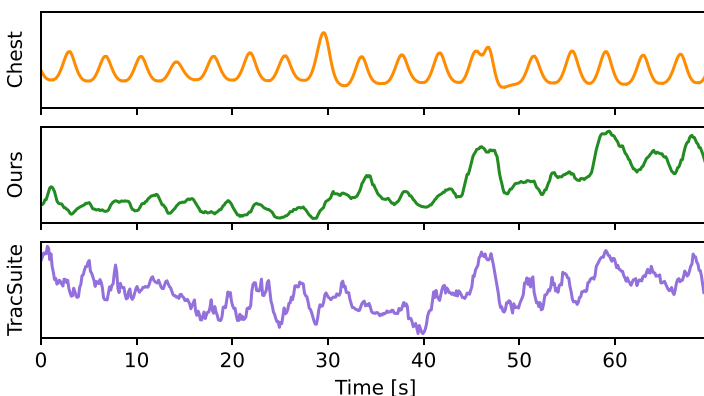
Table A1

We report the average motion for all MR scans in the protocol, as defined by our motion score (see Section 2.3). Additionally we show the pair-wise linear correlation coefficient of motion measured during the resting state fMRI sequence and other sequences. Finally, we show the approximate length of each image acquisition in seconds.

Sequence	average motion [mm/s]	Pearson's-r to fMRI	scan duration [s]
Resting state functional MRI	0.255 ± 0.126	1.000	625
T1-weighted imaging	0.273 ± 0.124	0.879	395
T2-weighted imaging	0.301 ± 0.162	0.863	287
Fluid-attenuated inversion recovery (FLAIR)	0.301 ± 0.153	0.824	277
Diffusion-weighted imaging	0.315 ± 0.147	0.768	682
Quantitative Susceptibility Mapping (QSM)	0.313 ± 0.148	0.775	380

Appendix B. Qualitative comparison of respiration- and head motion signals

We show a qualitative visualization of the respiration signal (relative chest movement), TracSuite, and our method (motion quantified by HPD



to initial reference depth image). Since TracSuite does not offer post-processing, we use our smooth equidistant re-sampling method to de-noise and re-sample both sequences equally. The result in Fig. B1 shows respiratory movements in both motion traces, however, the TracSuite signal contains a lot of "jitter", which could indicate noisy measurements. This is also reflected in the mutual information scores (TracSuite: 0.40, Ours: 0.73), even though the case was chosen to be as close as possible to both methods' 50th percentile mutual information score. At 45 seconds we can observe a distinct pattern of an elongated spike with two peaks in all three signals.

Appendix C. Induced motion experiment

We show the comparison of fMRI-based motion estimates to our registration method and TracSuite for one case, where the participant was asked to move in the scanner with removed padding. This results in much higher motion levels, than in the Rhineland Study (Fig. 4), causing a larger difference between "no registration" baseline and motion estimates of MCFLIRT. We use the fMRI frames to re-adjust our method for higher motion levels, as shown in Section 3.1, resulting in the parameters: reference point cloud = T1w MRI, regularizer weight = 0, max. optimizer steps = 50, $r_b = -0.5$, criterion function = Huber's function. For this single recorded high-motion case the difference of methods is not significant. Results can be seen in Fig. C1 below.

Transformation series distance to McFLIRT for a case of extreme motion

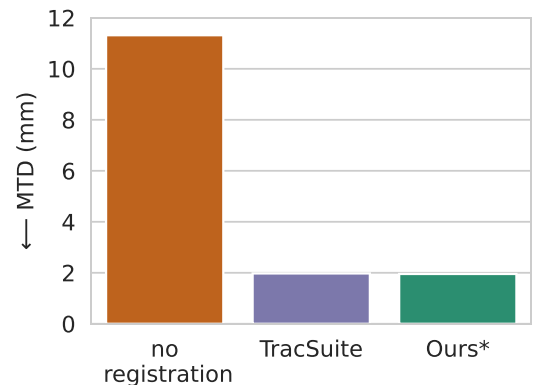


Fig. C1. Comparison of registration methods with fMRI motion traces as reference standard (as Fig. 4) for one case of induced motion. "Ours*" is our method with adjusted parameters for higher motion levels.

Fig. B1. This figure shows a qualitative comparison of respiratory chest movements (top) with our registration method (middle) and TracSuite (bottom) for a short, representative time interval of a single participant.

Appendix D. Motion trace difference

Here we describe the computation of motion trace differences in more detail. Prior to comparison we ensure with re-sampling and registration (see Section 2.3) that the n sampling time points of the two different motion traces agree and that they are defined with respect to the same aligned coordinate system (here from T1w MRI).

A motion trace consists of multiple head poses that can be described as homogeneous transformation matrices $M_i = P_{ref} \rightarrow P_i$ mapping a reference measurement (e.g. point cloud captured at the start of the session) P_{ref} to a measurement at time point i (P_i) of the trace. For the second motion trace, this reference is usually collected at a different time point and, thus, transformations are not directly comparable. This could be solved by selecting the same time point as the reference for both traces, e.g. time point one $P_{ref} = P_1$ and multiplying every transformation with the inverse transformation of the first time-point $P_1 \rightarrow P_{1...n} = M_{1...n} \times M_1^{-1}$. However, fixing a single reference time point in this fashion introduces bias into the motion trace. When we consider M_1 to consist of registration error ϵ and true head pose T , we see that calculating $P_1 \rightarrow P_{1...n} = M_{1...n} \times M_1^{-1} = M_{1...n} \times (T \times \epsilon)^{-1}$ propagates the error ϵ to every head pose.

To avoid emphasizing a specific time point with unknown head pose error we rephrase the motion-traces with respect to all possible time points. Consequently, we calculate the MTD, by averaging the HPD of all n head poses with all possible n time points as the reference $P_i \rightarrow P_j$ with $i, j \in 1..n$. This results in the term:

$$\text{MTD}(A_{1...n}, B_{1...n}) = \frac{\sum_{k=1}^n \sum_{l=1}^n \text{HPD}(A_l A_k^{-1}, B_l B_k^{-1}, h_c, h_r)}{n^2} \quad (\text{D.1})$$

to compare two motion traces $A_{1...n}, B_{1...n}$ (i.e. a series of homogeneous transformation matrices or head poses), with a head model of radius $h_r = 82.5$ mm and center h_c . The resulting metric is a performant, symmetric, and interpretable measure of the difference between two transformation series, that describes the average difference of head pose estimates in millimeter.

Appendix E. Ethics

The Rhineland Study is carried out in accordance with the recommendations of the International Council for Harmonisation's "Good Clinical Practice" standards (ICH-GCP). Written informed consent was obtained from all participants according to the Declaration of Helsinki. Approval was granted by the Ethics Committee of University Bonn.

The raw data visible in Figure 1 and 2 is specifically acquired for this work. The participant has given informed consent to inclusion of the motion tracking raw data in this publication.

References

- Afacan, O., Erem, B., Roby, D.P., Roth, N., Roth, A., Prabhu, S.P., Warfield, S.K., 2016. Evaluation of motion and its effect on brain magnetic resonance image quality in children. *Pediatr. Radiol.* 46 (12), 1728–1735. doi:10.1007/s00247-016-3677-9.
- Aksoy, M., Maclaren, J., Bammer, R., 2017. Prospective motion correction for 3D pseudo-continuous arterial spin labeling using an external optical tracking system. *Magn. Reson. Imag.* 39, 44–52. doi:10.1016/j.mri.2017.01.018.
- Alexander-Bloch, A., Clasen, L., Stockman, M., Ronan, L., Lalonde, F., Giedd, J., Raz-nahan, A., 2016. Subtle in-scanner motion biases automated measurement of brain anatomy from in vivo MRI. *Hum. Brain Mapp.* 37 (7), 2385–2397.
- Alhamud, A., Tisdall, M.D., Hess, A.T., Hasan, K.M., Meintjes, E.M., van der Kouwe, A.J., 2012. Volumetric navigators for real-time motion correction in diffusion tensor imaging. *Magn. Reson. Med.* 68 (4), 1097–1108.
- Andrews-Hanna, J.R., Snyder, A.Z., Vincent, J.L., Lustig, C., Head, D., Raichle, M.E., Buckner, R.L., 2007. Disruption of large-scale brain systems in advanced aging. *Neuron* 56 (5), 924–935.
- Andronesi, O.C., Bhattacharyya, P.K., Bogner, W., Choi, I.-Y., Hess, A.T., Lee, P., Meintjes, E.M., Tisdall, M.D., Zaitsev, M., van der Kouwe, A., 2021. Motion correction methods for MRS: experts' consensus recommendations. *NMR Biomed.* 34 (5), e4364.
- Arun, K.S., Huang, T.S., Blostein, S.D., 1987. Least-squares fitting of two 3-D point sets. *IEEE Trans. Pattern. Anal. Mach. Intell.* (5) 698–700.
- Atkinson, D., Hill, D.L., Stoyile, P.N., Summers, P.E., Keevil, S.F., 1997. Automatic correction of motion artifacts in magnetic resonance images using an entropy focus criterion. *IEEE Trans. Med. Imag.* 16 (6), 903–910.

- Barkovich, M.J., Li, Y., Desikan, R.S., Barkovich, A.J., Xu, D., 2019. Challenges in pediatric neuroimaging. *Neuroimage* 185, 793–801.
- Bergström, P., Edlund, O., 2014. Robust registration of point sets using iteratively reweighted least squares. *Comput. Optim. Appl.* 58 (3), 543–561. doi:10.1007/s10589-014-9643-2.
- Beyer, F., Kharabian Masouleh, S., Huntenburg, J.M., Lampe, L., Luck, T., Riedel-Heller, S.G., Loeffler, M., Schroeter, M.L., Stumvoll, M., Villringer, A., et al., 2017. Higher body mass index is associated with reduced posterior default mode connectivity in older adults. *Hum. Brain Mapp.* 38 (7), 3502–3515.
- Beyer, F., Prehn, K., Wüsten, K.A., Villringer, A., Ordemann, J., Flöel, A., Witte, A.V., 2020. Weight loss reduces head motion: Revisiting a major confound in neuroimaging. *Hum. Brain Mapp.* 41 (9), 2490–2494.
- Billar, A., Reuter, M., Patenaude, B., Homola, G., Breuer, F., Bendszus, M., Bartsch, A., 2015. Responses of the human brain to mild dehydration and rehydration explored in vivo by 1h-MR imaging and spectroscopy. *Am. J. Neuroradiol.* 36 (12), 2277–2284.
- Brenner, D., Stirnberg, R., Pracht, E.D., Stöcker, T., 2014. Two-dimensional accelerated mp-rage imaging with flexible linear reordering. *Magn. Reson. Mater. Phys., Biol. Med.* 27 (5), 455–462.
- Breteler, M.M., Stöcker, T., Pracht, E., Brenner, D., Stirnberg, R., 2014. MRI in the Rhineland study: A novel protocol for population neuroimaging. *Alzheimer's Dementia. J. Alzheimer's Assoc.* 10 (4), 92. doi:10.1016/j.jalz.2014.05.172.
- Callaghan, M.F., Josephs, O., Herbst, M., Zaitsev, M., Todd, N., Weiskopf, N., 2015. An evaluation of prospective motion correction (PMC) for high resolution quantitative MRI. *Front. Neurosci.* 9, 97.
- Castella, R., Arn, L., Dupuis, E., Callaghan, M.F., Draganski, B., Lutti, A., 2018. Controlling motion artefact levels in mr images by suspending data acquisition during periods of head motion. *Magn. Reson. Med.* 80 (6), 2415–2426.
- Chan, M.Y., Park, D.C., Savalia, N.K., Petersen, S.E., Wig, G.S., 2014. Decreased segregation of brain systems across the healthy adult lifespan. *Proc. Natl. Acad. Sci.* 111 (46), E4997–E5006.
- Costa, A.F., Petrie, D.W., Yen, Y.-F., Drangova, M., 2005. Using the axis of rotation of polar navigator echoes to rapidly measure 3D rigid-body motion. *Magn. Reson. Med.: Off. J. Int. Soc. Magn. Reson. Med.* 53 (1), 150–158.
- Couvy-Duchesne, B., Blokland, G.A., Hickie, I.B., Thompson, P.M., Martin, N.G., de Zubicaray, G.I., McMahon, K.L., Wright, M.J., 2014. Heritability of head motion during resting state functional MRI in 462 healthy twins. *Neuroimage* 102, 424–434.
- Couvy-Duchesne, B., Ebejer, J.L., Gillespie, N.A., Duffy, D.L., Hickie, I.B., Thompson, P.M., Martin, N.G., de Zubicaray, G.I., McMahon, K.L., Medland, S.E., et al., 2016. Head motion and inattention/hyperactivity share common genetic influences: implications for fMRI studies of ADHD. *PLoS ONE* 11 (1), e0146271.
- Cox, A.D., Virues-Ortega, J., Julio, F., Martin, T.L., 2017. Establishing motion control in children with autism and intellectual disability: applications for anatomical and functional MRI. *J. Appl. Behav. Anal.* 50 (1), 8–26.
- Dbouk, T., Drikakis, D., 2020. On coughing and airborne droplet transmission to humans. *Phys. Fluid.* 32 (5), 053310.
- Dosenbach, N.U., Koller, J.M., Earl, E.A., Miranda-Dominguez, O., Klein, R.L., Van, A.N., Snyder, A.Z., Nagel, B.J., Nigg, J.T., Nguyen, A.L., et al., 2017. Real-time motion analytics during brain MRI improve data quality and reduce costs. *Neuroimage* 161, 80–93.
- Eberhardt, O., Topka, H., 2017. Myoclonic disorders. *Brain Sci.* 7 (8), 103.
- Einspänner, E., Jochimsen, T.H., Harries, J., Melzer, A., Unger, M., Brown, R., Thiele-mans, K., Sabri, O., Sattler, B., 2022. Evaluating different methods of MR-based motion correction in simultaneous PET/MR using a head phantom moved by a robotic system. *EJNMMI Phys.* 9 (1), 1–17.
- Ekhtiari, H., Kupklicki, R., Yeh, H.-w., Paulus, M.P., 2019. Physical characteristics not psychological state or trait characteristics predict motion during resting state fMRI. *Sci. Rep.* 9 (1), 1–8.
- Engelhardt, L.E., Roe, M.A., Juranek, J., DeMaster, D., Harden, K.P., Tucker-Drob, E.M., Church, J.A., 2017. Children's head motion during fMRI tasks is heritable and stable over time. *Dev. Cogn. Neurosci.* 25, 58–68.
- Esteban, O., Birman, D., Schaer, M., Koyejo, O.O., Poldrack, R.A., Gorgolewski, K.J., 2017. MRIQC: advancing the automatic prediction of image quality in MRI from unseen sites. *PLoS ONE* 12 (9), e0184661.
- Fantini, I., Rittner, L., Yasuda, C., Lotufo, R., 2018. Automatic detection of motion artifacts on MRI using Deep CNN. In: *PRNI. IEEE*, pp. 1–4.
- Fischl, B., 2012. Freesurfer. *Neuroimage* 62 (2), 774–781.
- Forman, C., Aksoy, M., Hornegger, J., Bammer, R., 2011. Self-encoded marker for optical prospective head motion correction in MRI. *Med. Image Anal.* 15 (5), 708–719.
- Frew, S., Samara, A., Shearer, H., Eilbott, J., Vanderwal, T., 2022. Getting the nod: pediatric head motion in a transdiagnostic sample during movie-and resting-state fMRI. *PLoS ONE* 17 (4), e0265112.
- Frost, R., Wighton, P., Karahanoglu, F.I., Robertson, R.L., Grant, P.E., Fischl, B., Tisdall, M.D., van der Kouwe, A., 2019. Markerless high-frequency prospective motion correction for neuroanatomical MRI. *Magn. Reson. Med.* 82 (1), 126–144.
- Fu, Z.W., Wang, Y., Grimm, R.C., Rossman, P.J., Felmlee, J.P., Riederer, S.J., Ehman, R.L., 1995. Orbital navigator echoes for motion measurements in magnetic resonance imaging. *Magn. Reson. Med.* 34 (5), 746–753.
- Gilmore, A.D., Buser, N.J., Hanson, J.L., 2021. Variations in structural MRI quality significantly impact commonly used measures of brain anatomy. *Brain Inform.* 8 (1), 1–15.
- Goto, M., Abe, O., Miyati, T., Yamasue, H., Gomi, T., Takeda, T., 2015. Head motion and correction methods in resting-state functional MRI. *Magn. Reson. Med. Sci. rev.* 2015.
- Gramkow, C., 2001. On averaging rotations. *J. Math. Imag. Vis.* 15 (1), 7–16.
- Greene, D.J., Koller, J.M., Hampton, J.M., Wesевич, V., Van, A.N., Nguyen, A.L., Hoyt, C.R., McIntyre, L., Earl, E.A., Klein, R.L., et al., 2018. Behavioral interventions for reducing head motion during MRI scans in children. *Neuroimage* 171, 234–245.

- Greve, D.N., Fischl, B., 2009. Accurate and robust brain image alignment using boundary-based registration. *Neuroimage* 48 (1), 63–72.
- Henschel, L., Kügler, D., Reuter, M., 2022. FastSurferVINN: Building resolution-independence into deep learning segmentation methods - A solution for HighRes brain MRI. *Neuroimage* 251, 118933.
- Herbst, M., Lovell-Smith, C., Haeublein, B., Sostheim, R., Maclaren, J.R., Korvink, J.G., Zaitsev, M., 2013. On the robustness of prospective motion correction for clinical routine. In: Proceedings of the 21st Scientific Meeting: International Society for Magnetic Resonance in Medicine, p. 3766.
- Hess, A.T., Dylan Tisdall, M., Andronesi, O.C., Meintjes, E.M., van der Kouwe, A.J., 2011. Real-time motion and b0 corrected single voxel spectroscopy using volumetric navigators. *Magn. Reson. Med.* 66 (2), 314–323.
- Hodgson, K., Poldrack, R.A., Curran, J.E., Knowles, E.E., Mathias, S., Göring, H.H., Yao, N., Olvera, R.L., Fox, P.T., Almasy, L., et al., 2017. Shared genetic factors influence head motion during MRI and body mass index. *Cereb. Cortex* 27 (12), 5539–5546.
- Huijbers, W., Van Dijk, K.R., Boenniger, M.M., Stirnberg, R., Breteler, M.M., 2017. Less head motion during MRI under task than resting-state conditions. *Neuroimage* 147, 111–120. doi:10.1016/j.neuroimage.2016.12.002.
- Janos, S., Schooler, G.R., Ngo, J.S., Davis, J.T., 2019. Free-breathing unsaturated MRI in children: Justification and techniques. *J. Magn. Reson. Imaging* 50 (2), 365–376.
- Jenkinson, M., 1999. Measuring transformation error by RMS deviation. *Studholme, C., Hill, D.L.G., Hawkes, D.J.*
- Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* 17 (2), 825–841.
- Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M., 2012. FSL. *Neuroimage* 62 (2), 782–790.
- Kober, T., Marques, J.P., Gruetter, R., Krueger, G., 2011. Head motion detection using FID navigators. *Magn. Reson. Med.* 66 (1), 135–143.
- Kong, X.-z., Zhen, Z., Li, X., Lu, H.-h., Wang, R., Liu, L., He, Y., Zang, Y., Liu, J., 2014. Individual differences in impulsivity predict head motion during magnetic resonance imaging. *PLoS ONE* 9 (8), e104989.
- Van der Kouwe, A.J., Benner, T., Dale, A.M., 2006. Real-time rigid body motion correction and shimming using cloverleaf navigators. *Magn. Reson. Med.: Off. J. Int. Soc. Magn. Reson. Med.* 56 (5), 1019–1032.
- van der Kouwe, A.J., Benner, T., Salat, D.H., Fischl, B., 2008. Brain morphometry with multiecho MPRAGE. *Neuroimage* 40 (2), 559–569.
- Kraskov, A., Stögbauer, H., Grassberger, P., 2004. Estimating mutual information. *Phys. Rev. E* 69 (6), 066138.
- Küstner, T., Jandt, M., Liebgott, A., Mauch, L., Martirosian, P., Bamberg, F., Nikolaou, K., Giatidis, S., Schick, F., Yang, B., 2018. Automatic motion artifact detection for whole-body magnetic resonance imaging. In: ICASSP. IEEE, pp. 995–999.
- Küstner, T., Liebgott, A., Mauch, L., Martirosian, P., Bamberg, F., Nikolaou, K., Yang, B., Schick, F., Giatidis, S., 2018. Automated reference-free detection of motion artifacts in magnetic resonance images. *MAGMA* 31 (2), 243–256.
- Kyme, A.Z., Aksoy, M., Henry, D.L., Bammer, R., Maclaren, J., 2020. Marker-free optical stereo motion tracking for in-bore MRI and PET-MRI application. *Med. Phys.* 47 (8), 3321–3331.
- Largent, A., Kapse, K., Barnett, S.D., et al., 2021. Image quality assessment of fetal brain MRI using multi-instance deep learning methods. *JMRI* 54 (3), 818–829.
- Laustsen, M., Andersen, M., Xue, R., Madsen, K.H., Hanson, L.G., 2022. Tracking of rigid head motion during MRI using an EEG system. *Magn. Reson. Med.* 88 (2), 986–1001.
- Lei, K., Syed, A.B., Zhu, X., Pauly, J.M., Vasanawala, S.S., 2022. Artifact- and content-specific quality assessment for MRI with image rulers. *Med. Image Anal.* 77, 102344.
- Littleton, S.W., 2012. Impact of obesity on respiratory function. *Respirology* 17 (1), 43–49.
- Ma, J.J., Nakarmi, U., Kin, C.Y.S., Sandino, C.M., Cheng, J.Y., Syed, A.B., Wei, P., Pauly, J.M., Vasanawala, S.S., 2020. Diagnostic image quality assessment and classification in medical imaging: Opportunities and challenges. In: ISBI. IEEE, pp. 337–340.
- Maclaren, J., Armstrong, B.S., Barrows, R.T., Danishad, K., Ernst, T., Foster, C.L., Gummus, K., Herbst, M., Kadashevich, I.Y., Kusik, T.P., et al., 2012. Measurement and correction of microscopic head motion during magnetic resonance imaging of the brain. *PLoS ONE* 7 (11), e48088.
- Madan, C.R., 2018. Age differences in head motion and estimates of cortical morphology. *PeerJ* 6, e5176.
- Maknojia, S., Churchill, N.W., Schweizer, T.A., Graham, S., 2019. Resting state fMRI: Going through the motions. *Front. Neurosci.* 13, 825.
- Meissner, T.W., Walbrin, J., Nordt, M., Koldewyn, K., Weigelt, S., 2020. Head motion during fMRI tasks is reduced in children and adults if participants take breaks. *Dev. Cogn. Neurosci.* 44, 100803.
- Mortamet, B., Bernstein, M.A., Jack Jr, C.R., Gunter, J.L., Ward, C., Britson, P.J., Meuli, R., Thiran, J.-P., Krueger, G., 2009. Automatic quality assessment in structural brain magnetic resonance imaging. *Magn. Reson. Med.: Off. J. Int. Soc. Magn. Reson. Med.* 62 (2), 365–372.
- Mugler III, J.P., 2014. Optimized three-dimensional fast-spin-echo MRI. *J. Magn. Reson. Imaging* 39 (4), 745–767.
- Musa, M., Sengupta, S., Chen, Y., 2022. MRI-compatible soft robotic rensing pad for head motion detection. *IEEE Rob. Autom. Lett.* 7 (2), 3632–3639.
- Nakamura, K., Brown, R.A., Araujo, D., Narayanan, S., Arnold, D.L., 2014. Correlation between brain volume change and T2 relaxation time induced by dehydration and rehydration: Implications for monitoring atrophy in clinical studies. *NeuroImage: Clinical* 6, 166–170.
- Olesen, O.V., Jørgensen, M.R., Paulsen, R.R., Højgaard, L., Roed, B., Larsen, R., 2010. Structured light 3D tracking system for measuring motions in PET brain imaging. In: Medical Imaging 2010: Visualization, Image-Guided Procedures, and Modeling, Vol. 7625. SPIE, pp. 286–296.
- Olesen, O.V., Sullivan, J.M., Mulnix, T., Paulsen, R.R., Højgaard, L., Roed, B., Carson, R.E., Morris, E.D., Larsen, R., 2012. List-mode PET motion correction using markerless head tracking: Proof-of-concept with scans of human subject. *IEEE Trans. Med. Imaging* 32 (2), 200–209.
- Oztek, M.A., Noda, S., Beauchemin, E.A., Otto, R.K., 2020. Gentle touch: Noninvasive approaches to improve patient comfort and cooperation for pediatric imaging. *Top. Magn. Reson. Imag.* 29 (4), 187–195.
- Pannetier, N.A., Stavrinou, T., Ng, P., Herbst, M., Zaitsev, M., Young, K., Matson, G., Schuff, N., 2016. Quantitative framework for prospective motion correction evaluation. *Magn. Reson. Med.* 75 (2), 810–816.
- Pardoe, H.R., Hiess, R.K., Kuzniecky, R., 2016. Motion and morphometry in clinical and nonclinical populations. *Neuroimage* 135, 177–185.
- Pardoe, H.R., Martin, S.P., Zhao, Y., George, A., Yuan, H., Zhou, J., Liu, W., Devinsky, O., 2021. Estimation of in-scanner head pose changes during structural MRI using a convolutional neural network trained on eye tracker video. *Magn. Reson. Imaging* 81, 101–108.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.
- Pollak, C., Kügler, D., Reuter, M., 2023. Estimating head motion from MRI. In: Proceedings IEEE 20th International Symposium on Biomedical Imaging (ISBI). IEEE.
- Provins, C., Schöttner, M., Dayan, M., Nastase, V., Lunde, J., Benach, O.M., MacNicol, E., Savary, E., Seeley, S.H., Hagen, M.P., et al., 2023. Signal-to-noise ratio estimates predict head motion presence in T1-weighted MRI. *OSF Preprint* doi:10.31219/osf.io/7vqzr.
- Reuter, M., Rosas, H.D., Fischl, B., 2010. Highly accurate inverse consistent registration: a robust approach. *Neuroimage* 53 (4), 1181–1196. doi:10.1016/j.neuroimage.2010.07.020.
- Reuter, M., Schmansky, N.J., Rosas, H.D., Fischl, B., 2012. Within-subject template estimation for unbiased longitudinal image analysis. *Neuroimage* 61 (4), 1402–1418. doi:10.1016/j.neuroimage.2012.02.084.
- Reuter, M., Tisdall, M.D., Qureshi, A., Buckner, R.L., van der Kouwe, A.J., Fischl, B., 2015. Head motion during MRI acquisition reduces gray matter volume and thickness estimates. *Neuroimage* 107, 107–115. doi:10.1016/j.neuroimage.2014.12.006.
- Rizk-Jackson, A., Stoffers, D., Sheldon, S., Kuperman, J., Dale, A., Goldstein, J., Corey-Bloom, J., Poldrack, R.A., Aron, A.R., 2011. Evaluating imaging biomarkers for neurodegeneration in pre-symptomatic Huntington's disease using machine learning techniques. *Neuroimage* 56 (2), 788–796.
- Rosen, A.F., Roalf, D.R., Ruparel, K., Blake, J., Seelaus, K., Villa, L.P., Ciric, R., Cook, P.A., Davatzikos, C., Elliott, M.A., et al., 2018. Quantitative assessment of structural image quality. *Neuroimage* 169, 407–418.
- Savalia, N.K., Agres, P.F., Chan, M.Y., Feczko, E.J., Kennedy, K.M., Wig, G.S., 2017. Motion-related artifacts in structural brain images revealed with independent estimates of in-scanner head motion. *Hum. Brain Mapp.* 38 (1), 472–492.
- Sciarra, A., Chatterjee, S., Dünwald, M., Placidi, G., Nürnberger, A., Speck, O., Oeltze-Jaffra, S., 2022. Reference-less SSIM regression for detection and quantification of motion artefacts in brain MRIs. *MIDL*.
- Seabold, S., Perktold, J., 2010. Statsmodels: Econometric and statistical modeling with Python. In: 9th Python in Science Conference.
- Shehzad, Z., Givassios, S., Li, Q., Benhajali, Y., Yan, C., Yang, Z., Milham, M., Bellec, P., Craddock, C., 2015. The preprocessed connectomes project quality assessment protocol - A resource for measuring the quality of MRI data. *Front. Neurosci.* 47.
- Slipsager, J.M., Ellegaard, A.H., Glimberg, S.L., Paulsen, R.R., Tisdall, M.D., Wighton, P., van der Kouwe, A., Marner, L., Henriksen, O.M., Law, I., et al., 2019. Markerless motion tracking and correction for PET, MRI, and simultaneous PET/MRI. *PLoS ONE* 14 (4), e0215524.
- Slipsager, J.M., Glimberg, S.L., Højgaard, L., Paulsen, R.R., Wighton, P., Tisdall, M.D., Jaimes, C., Gagoski, B.A., Grant, P.E., van Der Kouwe, A., et al., 2022. Comparison of prospective and retrospective motion correction in 3D-encoded neuroanatomical MRI. *Magn. Reson. Med.* 87 (2), 629–645.
- Slipsager, J.M., Glimberg, S.L., Søgaard, J., Paulsen, R.R., Johannesen, H.H., Martens, P.C., Seth, A., Marner, L., Henriksen, O.M., Olesen, O.V., et al., 2020. Quantifying the Financial Savings of Motion Correction in Brain MRI: A Model-Based Estimate of the Costs Arising From Patient Head Motion and Potential Savings From Implementation of Motion Correction. *J. Magn. Reson. Imag.* 52 (3), 731–738.
- Stirnberg, R., Huijbers, W., Brenner, D., Poser, B.A., Breteler, M., Stöcker, T., 2017. Rapid whole-brain resting-state fMRI at 3 T: Efficiency-optimized three-dimensional EPI versus repetition time-matched simultaneous-multi-slice EPI. *Neuroimage* 163, 81–92. doi:10.1016/j.neuroimage.2017.08.031.
- Stöcker, T., 2016. Big data: The Rhineland study. In: Proceedings of the 24th Scientific Meeting of the International Society for Magnetic Resonance in Medicine.
- Stpień, I., Obuchowicz, R., Piórkowski, A., Oszust, M., 2021. Fusion of deep convolutional neural networks for no-reference magnetic resonance image quality assessment. *Sensors* 21 (4), 1043.
- Streitbürger, D.-P., Möller, H.E., Tittgemeyer, M., Hund-Georgiadis, M., Schroeter, M.L., Mueller, K., 2012. Investigating structural brain changes of dehydration using voxel-based morphometry. *PLoS One* 7 (8), e44195.
- Stucht, D., Danishad, K.A., Schulze, P., Godenschwager, F., Zaitsev, M., Speck, O., 2015. Highest resolution in vivo human brain MRI using prospective motion correction. *PLoS ONE* 10 (7), e0133921.
- Sujit, S.J., Coronado, I., Kamali, A., Narayana, P.A., Gabr, R.E., 2019. Automated image quality evaluation of structural brain MRI using an ensemble of deep learning networks. *J. Magn. Reson. Imag.* 50 (4), 1260–1267.
- Thenganatt, M.A., Jankovic, J., 2016. The relationship between essential tremor and Parkinson's disease. *Parkinsonism Rel. Disord.* 22, S162–S165.

- Tisdall, M.D., Hess, A.T., Reuter, M., Meintjes, E.M., Fischl, B., van der Kouwe, A.J., 2012. Volumetric navigators for prospective motion correction and selective reacquisition in neuroanatomical MRI. *Magn. Reson. Med.* 68 (2), 389–399. doi:10.1002/mrm.23228.
- Todd, N., Josephs, O., Callaghan, M.F., Lutti, A., Weiskopf, N., 2015. Prospective motion correction of 3D echo-planar imaging data for functional MRI using optical tracking. *Neuroimage* 113, 1–12.
- Tomasi, D., Volkow, N.D., 2013. Striatocortical pathway dysfunction in addiction and obesity: Differences and similarities. *Crit. Rev. Biochem. Mol. Biol.* 48 (1), 1–19.
- Torres, E.B., Denisova, K., 2016. Motor noise is rich signal in autism research and pharmacological treatments. *Sci. Rep.* 6 (1), 1–19.
- Vanderwal, T., Kelly, C., Eilbott, J., Mayes, L.C., Castellanos, F.X., 2015. Inscapes: A movie paradigm to improve compliance in functional magnetic resonance imaging. *Neuroimage* 122, 222–232.
- Versluis, M., Peeters, J., van Rooden, S., van der Grond, J., van Buchem, M.A., Webb, A.G., van Osch, M.J., 2010. Origin and reduction of motion and f0 artifacts in high resolution T2*-weighted magnetic resonance imaging: Application in Alzheimer's disease patients. *Neuroimage* 51 (3), 1082–1088.
- Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al., 2020. Scipy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* 17 (3), 261–272.
- Welch, E.B., Manduca, A., Grimm, R.C., Ward, H.A., Jack Jr, C.R., 2002. Spherical navigator echoes for full 3D rigid body motion measurement in MRI. *Magn. Reson. Med.: off. J. Int. Soc. Magn. Reson. Med.* 47 (1), 32–41.
- Wilke, M., 2012. An alternative approach towards assessing and accounting for individual motion in fMRI timeseries. *Neuroimage* 59 (3), 2062–2072.
- Wilke, M., 2014. Isolated assessment of translation or rotation severely underestimates the effects of subject motion in fMRI data. *PLoS ONE* 9 (10), e106498.
- Wylie, G.R., Genova, H., DeLuca, J., Chiaravalloti, N., Sumowski, J.F., 2014. Functional magnetic resonance imaging movers and shakers: Does subject-movement cause sampling bias? *Hum. Brain Mapp.* 35 (1), 1–13.
- Zaca, D., Hasson, U., Minati, L., Jovicich, J., 2018. Method for retrospective estimation of natural head movement during structural MRI. *J. Magn. Reson. Imag.* 48 (4), 927–937.
- Zaitsev, M., Maclaren, J., Herbst, M., 2015. Motion artifacts in MRI: A complex problem with many partial solutions. *J. Magn. Reson. Imag.* 42 (4), 887–901.
- Zeng, L.-L., Wang, D., Fox, M.D., Sabuncu, M., Hu, D., Ge, M., Buckner, R.L., Liu, H., 2014. Neurobiological basis of head motion in brain imaging. *Proc. Natl. Acad. Sci.* 111 (16), 6058–6062.
- Zukić, D., Haley, A., Lisle, C., Klo, J., Pohl, K.M., Johnson, H.J., Chaudhary, A., 2022. Medical image quality assurance using deep learning. *MIDL*.

Appendix B.

Publication: "Estimating Head Motion from Mr-Images"

This is an author manuscript distributed under the terms of the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. ©2023 The Authors.

Estimating head motion from MR-images.

Authors: **Clemens Pollak**, David Kügler, and Martin Reuter.

Published in: arXiv.org preprint

DOI: [10.48550/arXiv.2302.14490](https://doi.org/10.48550/arXiv.2302.14490)

Github: [Deep-MI/head-motion-from-MRI](https://github.com/Deep-MI/head-motion-from-MRI)

The peer-reviewed article is available at the digital object identifier (DOI):
[10.1109/ISBI53787.2023.10230717](https://doi.org/10.1109/ISBI53787.2023.10230717) ©2023 IEEE.

Estimating head motion from MR-images.

Authors: **Clemens Pollak**, David Kügler, and Martin Reuter.

Published in: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), pp. 1-5. IEEE

DOI: [10.1109/ISBI53787.2023.10230717](https://doi.org/10.1109/ISBI53787.2023.10230717)

Github: [Deep-MI/head-motion-from-MRI](https://github.com/Deep-MI/head-motion-from-MRI)

ESTIMATING HEAD MOTION FROM MR-IMAGES

Clemens Pollak^{1,*}, David Kügler^{1,*}, and Martin Reuter^{1,2,3}

¹AI in Medical Imaging, German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

²A.A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston, MA, USA

³Department of Radiology, Harvard Medical School, Boston, MA, USA

ABSTRACT

Head motion is an omnipresent confounder of magnetic resonance image (MRI) analyses as it systematically affects morphometric measurements, even when visual quality control is performed. In order to estimate subtle head motion, that remains undetected by experts, we introduce a deep learning method to predict in-scanner head motion directly from T1-weighted (T1w), T2-weighted (T2w) and fluid-attenuated inversion recovery (FLAIR) images using motion estimates from an in-scanner depth camera as ground truth. Since we work with data from compliant healthy participants of the Rhineland Study, head motion and resulting imaging artifacts are less prevalent than in most clinical cohorts and more difficult to detect. Our method demonstrates improved performance compared to state-of-the-art motion estimation methods and can quantify drift and respiration movement independently. Finally, on unseen data, our predictions preserve the known, significant correlation with age.

Keywords Motion estimation · MRI quality · Deep learning · Motion tracking

Corresponding author: Martin Reuter (martin.reuter[at]dzne.de)

1 Introduction

Head motion is a ubiquitous challenge for magnetic resonance image (MRI) acquisition. It causes a range of image artifacts that introduce bias in downstream analysis [1–6], which persists despite expert quality control [1, 2]. While initially explored for clinical cohorts with increased motion levels [5–8] or induced motion [1, 9], less research focuses on motion in studies of healthy, compliant population cohorts [2, 10] such as the Rhineland Study [11, 12]. Critically, the lack of a sensitive and reliable motion estimation method to quantify subtle motion hinders the inclusion of motion estimates in statistical models to control motion-induced biases. For example, careful visual inspection of the Rhineland

Study dataset, used in this paper, did not detect any cases with clearly visible motion artefacts that would warrant exclusion. Yet, even in the 75-participant subset reserved for testing, a statistically significant correlation of motion with age can be shown, underlining the need for sensitive estimation and control of head motion in MRI analyses.

In this paper, we propose a method to directly estimate head motion from the acquired MR image. We measure head motion during MRI acquisition via head tracking with a depth camera and establish a ground truth motion score per sequence. This is contrary to the currently established paradigm of predicting discrete motion severity levels established via an expert manual quality control process [8–10, 13–18]. Expert ratings are limited by their subjectivity to

*These authors contributed equally to this work

the specific task and human perception [18–20] hindering their general utility, specifically for compliant, low-motion cohorts. Camera-based motion measurements, on the contrary, are objective and sensitive even for low-motion cohorts, where motion-induced image artifacts are almost invisible.

Since the rise of deep learning, tools have been able to predict expert motion ratings with increasing accuracy [8–10, 13–18] sometimes addressing previous limitations, for example the subjectivity to the task [18]. Currently, the only alternative to the prediction of expert labels is to predict a perceptual image similarity metric between low motion “baseline” images and high-motion images, which are retrospectively simulated [21]. This approach replaces the human annotation task by a comparison of images with a perceptual similarity metric (SSIM), which may also suffer from similar limitations as expert labels. Moreover, the methods accuracy on real-world data relies on realistic, high-quality simulation, which also has to be adapted to each acquisition sequence. Meanwhile, our method can be directly re-trained even on different modalities, without any changes. Recent work in the field of in-MRI motion tracking enabled highly accurate tracking of head-motion during acquisition with the MRI scanner [22, 23], optical cameras [24, 25] or other devices [26, 27]. Yet, until tracking devices and methods are deployed to all imaging sites, image-derived motion estimates may help reduce potential motion induced biases – even retrospectively.

Our contributions are threefold – we 1. introduce – for the first time – the estimation of an objective motion score from images of three MRI sequences, 2. present a deep-learning-based solution, which outperforms DenseNet and state-of-the-art quality/motion estimation methods on a dataset of compliant, low-motion participants, and 3. quantify motion from respiration and (relaxation) drift. Finally, our method detects the significant, known correlation between predicted motion and age. We will publish our code on GitHub*.

2 Materials & Methods

2.1 Data acquisition

The Rhineland study is an ongoing population study recruiting a representative cohort of healthy participants above the age of 30. Our dataset describes a subset (ages 30 to 95 years) of 500 participants (282 female) with T1w, T2w and FLAIR images (at 0.8, 0.8 and 1.0 mm isotropic voxel size,

*<https://github.com/Deep-MI/head-motion-from-MRI/>

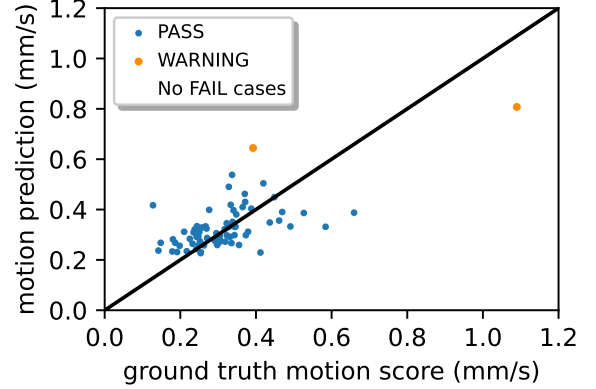


Figure 1: Plot of ground truth and predicted motion score. Blue *PASS* and orange *WARN* (visible artifacts) test images may be perfectly separated with a threshold of 0.6.

respectively) following a standardized acquisition protocol*. This dataset also includes expert quality-labels of T1w images (*PASS*, *WARN* or *FAIL*), where *WARN* indicates visible and *FAIL* strong artifacts (insufficient for downstream analysis). However, no images are rated as *FAIL* and only 9 images as *WARN*. The low number of *WARN* and *FAIL* cases is likely founded on high compliance and extensive efforts to reduce head motion during MRI acquisition including tight head padding, scheduled speaking breaks, careful participant instruction, and calming nature scenes shown during the scan.

To quantify the participant’s head motion, a video of depth images showing a portion of their face is collected concurrently to the scan [24]. Individual frames are aligned with a reference frame resulting in time series of rigid transformations. To obtain a per-sequence, scalar *motion score*, we 1. synchronize MRI and depth camera, 2. compute Jenkinson’s transformation differences [29] per pair of transformations, and 3. extract and average values for the duration of individual sequences. The Jenkinson’s transformation difference summarizes rigid transformations by averaging displacements within a spherical head model. The final *motion score* quantifies the average motion in millimeter per second. We randomly split the dataset into training, validation, and evaluation sets of 350, 75 and 75 participants, respectively.

2.2 Distinguishing motion patterns

In addition to the per-sequence motion score, we distinguish between three prominent types of in-scanner head motion

*For details on the acquisition of 3D T1w MPRAGE (scan length 6.5 min), 3D T2w (4.6 min) and FLAIR (4.5 min) images, see Lohner et al. [28].

corresponding to three frequency bands: i) head drift, ii) periodic motion due to breathing, and iii) “noisy motion”, which we expect is hard to estimate. To determine global filter thresholds, we estimate upper and lower median respiratory frequency of participants during the T1w sequence from an independent respiration sensor as 0.1 Hz and 0.5 Hz. We apply symmetric Butterworth filters (low-pass, band-pass and high-pass, respectively) to the time series of Jenkinson’s transformation differences. From the filtered signals, we aggregate the motion score as before resulting in three frequency-dependent targets.

2.3 Neural network & training

For the estimation of motion scores from 3D MR images, we adopt a fully convolutional neural network (CNN) from brain age estimation [30], an established regression task in medical imaging. The lightweight architecture permits training the 3D CNN on a single NVIDIA A100 GPU with batch size two. Instead of directly regressing the motion score, we follow Peng et al. [30] in their approach and 1. generously define the expected range of motion [0, 3.12] mm/s, 2. split it into 40 bins of ‘prototypes’, 3. for each prototype calculate the probability that the current motion score belongs into the prototype, and 4. train the CNN using a Kullback-Leibler loss (Adam optimizer for 500 epochs, approx. 10 h on one A100 GPU). To reconstruct the motion score from the predicted probability distribution, we sum the product of prototype centers and predictions. Many standard data augmentation strategies are not suitable for motion estimation, since the re-sampling of images affects the image noise, which is why we avoid interpolation of images completely. Helpful data augmentation, on the other hand, include intensity scaling in the range [0.9, 1.1] and random flipping with 30% probability along all axes. In our ablation study (Section 3.3), we explore different pre-processing operations. A focus on the 8 Least Significant Bits (LSB8) is useful for this task, leaving only an integer representation of the fine image differences.

2.4 Evaluation & statistical methods

We evaluate the regression model with the coefficient of determination (R^2 score – a measure for the average error) and Spearman’s rank correlation coefficient (Spearman’s ρ – a measure for correct ranking). The R^2 score normalizes the mean squared error to a range of $[-\infty, 1]$, where score < 0 indicates a prediction error worse than a constant prediction of the dataset mean and a score of 1 indicates perfect predictions. The Spearman’s ρ , on the other hand, is defined in

	Method	R^2	Spr- ρ
re-trained	Ours	0.433	0.584
	DenseNet [31]	0.395	0.447
	SFCN [30]	0.275	0.454
	MIQA CNN [8]	-0.273	0.192
QCtools	MIQA MA / QS	-	0.243 / 0.240
	MIQA MA / QS ¹	-	- ² / 0.338
	AES [7]	-	0.110

Table 1: Our methods outperforms SOTA approaches in both Spearman’s ρ (Spr- ρ) and R^2 scores (only valid for predicted motion scores) when predicting motion scores from T1w images on the test set. MA: motion artefact score, QS: quality score, ¹images standardized, ²no motion detected

the range [0, 1]. It is not affected by large, absolute errors of outliers and more sensitive to prediction errors, where the sampling of values is denser (i.e. more sensitive to small errors on values close together). We also analyze the rank correlation between motion and age using this method. We use the R^2 score as the primary metric for ablation, and select parameters that have the highest R^2 score on the validation set in experiments.

3 Results

We visualize the performance of our method on the unseen test set in Figure 1, which illustrates good correlation between ground truth measured motion (horizontal axis) and predictions from images (vertical). Perfect predictions would lie on the black line. Our method perfectly separates *PASS* (no artifacts) and *WARN* (mild artifacts) cases. A horizontal separation line at ≈ 0.6 mm/s can be found but no vertical line for the ground truth motion score.

3.1 Comparison with state-of-the-art motion estimation

To the best of our knowledge, there is currently no competing method to predict measured, in-scanner head motion from MR images. Since quality estimation methods cannot be easily re-trained on our dataset, which has few *WARN* and no *FAIL* labels, we compare with the pre-trained MIQA quality estimator [8] and Average Edge Strength (AES) [7], a heuristic known to correlate with motion [7]. Additionally, we compare our method with three deep learning architectures re-trained on our dataset: i) DenseNet [31], ii) SFCN [30], a CNN for brain age prediction, and iii) the CNN used by MIQA [8]. Our method, which is an improvement of SFCN (e.g. initializer), achieves the best correlation

Input	Target	R^2	Spr- ρ
T1	motion score	0.433	0.584
T2	motion score	0.362	0.556
FLAIR	motion score	0.299	0.489
T1	drift	0.183	0.637
T1	breathing band	0.185	0.382
T1	noisy motion	0.050	0.337

Table 2: The evaluations show generalization of our method to predicting the motion score on T2w and FLAIR images and very promising Spr- ρ performance for detection of drift on the test set.

with ground truth motion scores on the unseen test set (Table 1). The negative R^2 of the re-trained MIQA CNN indicates failed generalization.

The pre-trained MIQA tool aggregates predictions of expert ratings for 9 artifact types including a ‘motion artifacts’ (MA) score into a continuous quality score (QS). We select their probabilities as potential correlates of our motion score. Since MIQA has been trained on the 1 mm T1w-PREDICT-HD dataset, we test, whether rescaling and resampling images with FastSurfer’s [32] conform tool to 1 mm reduces domain shift. We measure a low but significant correlation between QS and motion score as well as between MA and motion score on the native 0.8 mm images. While previous work reported that probabilities of ratings, like those of MIQA, quantify subtle differences in motion, despite binary ground truth labels [16], MA probabilities are zero for all conformed images (consistent with the lack of manual *FAIL* labels). QS, on the other hand, significantly correlates with the motion score. In addition to the deep learning estimators, we evaluate the performance of AES [7], but do not find significant correlation between AES and the ground truth motion score.

3.2 Generalization to T2/FLAIR and motion types

We test the generalizability of our method to new tasks by predicting the motion score for T2w and FLAIR images, as well as predicting drift and respiratory motion on T1w images. For each task we re-train the network architecture and show the results in Table 2. While the R^2 score is not directly comparable across different tasks, we find good performance by purely re-training the model for these modalities. The Spearman’s- ρ in T2w and FLAIR experiments is similar to T1w experiments despite optimization on T1w only.

To quantify different motion types, we define two distinct aggregates of participant motion: i) slow relaxation-drift over

Method	Images	R^2	Spr- ρ
CNN MSE loss	T1 (LSB8)	0.117	0.376
CNN ours	T1 (unprocessed)	0.341	0.551
CNN ours	T1 (robust scaling)	0.322	0.489
CNN ours	T1 (remove head)	0.369	0.525
CNN ours	T1 (LSB8)	0.393	0.491

Table 3: Ablation of loss function and image pre-processing on the validation set. LSB8: 8 Least Significant Bits.

time, and ii) periodic head motion due to breathing. We filter motion estimates as described in Section 2.2 and train our architecture to predict the aggregates (Table 2). Our method can predict both slow drift (low frequency) motion and respiratory (medium frequency) motion from the T1w images. The aggregate of high frequencies, which are not associated with known motion types (noisy motion), cannot be predicted.

3.3 Ablation study

To optimize the parameters for our method, we compare choices for loss and preprocessing on the validation set. A critical finding of this work is that direct prediction trained using an Mean Squared Error (MSE) loss only achieves mild correlation with the motion score, while using a multi-dimensional probability distribution together with a Kullback-Leibler divergence loss yields a large performance uplift. For this MSE-loss ablation in Table 3 (top), we reduce the last layer to a single output and remove the softmax.

Several measures of image quality have taken advantage of the background signal to determine a reduction in image quality [33, 34]. Consequently, we explore the effects of four image pre-processing operations on the prediction quality in Table 3: No pre-processing, FastSurfer’s [32] robust scaling (removing 20% of high intensity voxels), removing the head with FreeSurfer’s head segmentation tool [35, 36] (the result is just the background) and dropping the Most Significant Bits leaving us with the 8 Least Significant Bits (LSB8). The latter approach, which drops the information about the absolute size of values directly on the images integer representation, surprisingly outperforms other ways of adjusting image intensities. This finding was validated in multiple experiments and across additional hypotheses.

3.4 Correlation with age

In adult populations, increased head motion in the MR scanner is associated with the increased age of participants [4,

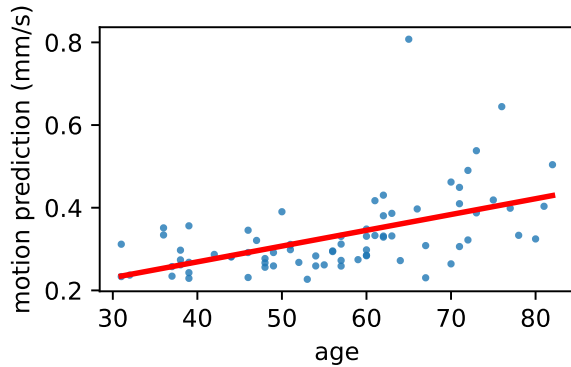


Figure 2: Age is significantly correlated with our motion predictions. This reflects the ground truth motion score labels (not shown) and is outlined by a linear fit, displayed as a red line.

37]. We can also measure this correlation within the ground truth motion score using Spearman’s ρ and find a significant correlation with age on the whole dataset as well as only on the test set ($p < 0.001$). Our predictions also present this correlation with age on the test set ($p < 0.001$) as illustrated in Figure 2 with the linear fit.

4 Discussion & conclusion

We introduce a novel task of in-scanner head motion estimation directly from MR images. The presented deep learning method provides sensitive predictions of motion levels capturing subtle correlations with known confounders such as age – all in unseen images that pass visual inspection. Two factors contribute to this sensitivity: predicting the probability distribution of ‘prototypes’ together with the Kullback-Leibler divergence loss, and the input of least-significant-bit (LSB) images. Why LSB images are preferable to raw structural images or their background should be further investigated in future work.

A well-known limitation of deep learning methods is the limited generalizability to unseen datasets. Large differences in motion levels between cohorts and the chosen acquisition parameters greatly affect the appearance of motion artifacts, hence we expect dedicated training datasets will be required to for re-training and generalization to unseen MR imaging sequences. Additionally, future work should investigate, whether motion estimation itself is also affected by biases, like age and diseases, which are known to be an indicator of increased motion levels.

Our method ranks images by their motion level better than comparable, state-of-the-art methods for MRI motion and quality estimation from expert ratings. Therefore, it may aid

quality control procedures in the identification and exclusion of cases with artifacts, which is also indicated by the clear separation between *PASS* and *WARN* labels in our experiments (Figure 1). However, confirmation on a dataset with more strongly motion-affected cases is required. Additionally, our method transfers well to the prediction of alternative targets for respiratory- and drift motion and from T2w and FLAIR images with comparable accuracy.

Finally, our method enables an analysis of other, perhaps unknown, correlates of motion as well as the integration of motion scores as a control variable in statistical models. This is particularly valuable in longitudinal cohort studies, like the Rhineland Study, to disentangle the bias of motion effects from other effects such as participant’s age and diseases.

5 Compliance with ethical standards

The Rhineland Study is carried out in accordance with the recommendations of the ICH-GCP standards. Written informed consent was obtained from all participants in accordance with the Declaration of Helsinki. Approval was granted by the Ethics Committee of University Bonn.

6 Acknowledgments

This work was supported by DZNE institutional funds, by the Federal Ministry of Education and Research of Germany (031L0206), and the Helmholtz-AI project DeGen (ZT-I-PF-5-078). We thank the Rhineland Study group (PI Monique Breteler) for supporting the data acquisition and management. The authors do not have any conflict of interest.

References

- [1] M. Reuter et al. “Head motion during MRI acquisition reduces gray matter volume and thickness estimates”. In: *Neuroimage* 107 (2015), pp. 107–115.
- [2] A. Alexander-Bloch et al. “Subtle in-scanner motion biases automated measurement of brain anatomy from in vivo MRI”. In: *HBM* 37.7 (2016), pp. 2385–2397.
- [3] A. D. Gilmore, N. J. Buser, and J. L. Hanson. “Variations in structural MRI quality significantly impact commonly used measures of brain anatomy”. In: *Brain informatics* 8.1 (2021), pp. 1–15.
- [4] N. K. Savalia et al. “Motion-related artifacts in structural brain images revealed with independent estimates of in-scanner head motion”. In: *HBM* 38.1 (2017), pp. 472–492.

- [5] H. R. Pardoe, R. K. Hiess, and R. Kuzniecky. "Motion and morphometry in clinical and nonclinical populations". In: *Neuroimage* 135 (2016), pp. 177–185.
- [6] A. F. Rosen et al. "Quantitative assessment of structural image quality". In: *Neuroimage* 169 (2018), pp. 407–418.
- [7] Domenico Zaca et al. "Method for retrospective estimation of natural head movement during structural MRI". In: *Journal of Magnetic Resonance Imaging* 48.4 (2018), pp. 927–937.
- [8] D. Zukić et al. "Medical Image Quality Assurance using Deep Learning". In: *MIDL*. 2022.
- [9] T. Küstner et al. "Automatic motion artifact detection for Whole-Body magnetic resonance imaging". In: *ICASSP*. IEEE. 2018, pp. 995–999.
- [10] I. Fantini et al. "Automatic detection of motion artifacts on MRI using Deep CNN". In: *PRNI*. IEEE. 2018, pp. 1–4.
- [11] M. Breteler et al. "Mri in the rhineland study: a novel protocol for population neuroimaging." In: *Alzheimer's Dement.* 10 (2014), p. 92.
- [12] T. Stöcker. "Big data: the Rhineland study". In: *ISMRM*. 2016.
- [13] A. Largent, K. Kapse, S. D. Barnett, et al. "Image quality assessment of fetal brain MRI using multi-instance deep learning methods". In: *JMRI* 54.3 (2021), pp. 818–829.
- [14] S. J. Sujit et al. "Automated image quality evaluation of structural brain MRI using an ensemble of deep learning networks". In: *JMRI* 50.4 (2019), pp. 1260–1267.
- [15] J. J. Ma et al. "Diagnostic image quality assessment and classification in medical imaging: Opportunities and challenges". In: *ISBI*. IEEE. 2020, pp. 337–340.
- [16] T. Küstner et al. "Automated reference-free detection of motion artifacts in magnetic resonance images". In: *MAGMA* 31.2 (2018), pp. 243–256.
- [17] I. Stpień et al. "Fusion of Deep Convolutional Neural Networks for No-Reference Magnetic Resonance Image Quality Assessment". In: *Sensors* 21.4 (2021), p. 1043.
- [18] K. Lei et al. "Artifact-and content-specific quality assessment for MRI with image rulers". In: *Medical Image Analysis* 77 (2022), p. 102344.
- [19] H. H. Barrett et al. "Task-based measures of image quality and their relation to radiation dose and patient risk". In: *Physics in Medicine & Biology* 60.2 (2015), R1.
- [20] M. Pryde et al. "The Performance of Image Quality Metrics Depends on the Diagnostic Task: A Case Study in Stroke MRI". In: *ISMRM* (2022).
- [21] A. Sciarra et al. "Reference-less SSIM Regression for Detection and Quantification of Motion Artefacts in Brain MRIs". In: *MIDL*. 2022.
- [22] M. D. Tisdall et al. "Volumetric navigators for prospective motion correction and selective reacquisition in neuroanatomical MRI". In: *Magn. Reson. Med.* 68.2 (2012), pp. 389–399.
- [23] O. C. Andronesi, P. K. Bhattacharyya, W. Bogner, et al. "Motion correction methods for MRS: experts' consensus recommendations". In: *NMR in Biomedicine* 34.5 (2021), e4364.
- [24] J. M. Slipsager, A. H. Ellegaard, S. L. Glimberg, et al. "Markerless motion tracking and correction for PET, MRI, and simultaneous PET/MRI". In: *Plos one* 14.4 (2019), e0215524.
- [25] H. R. Pardoe et al. "Estimation of in-scanner head pose changes during structural MRI using a convolutional neural network trained on eye tracker video". In: *Magnetic Resonance Imaging* 81 (2021), pp. 101–108.
- [26] M. Musa, S. Sengupta, and Y. Chen. "MRI-Compatible Soft Robotic Sensing Pad for Head Motion Detection". In: *IEEE Robotics and Automation Letters* 7.2 (2022), pp. 3632–3639.
- [27] M. Laustsen et al. "Tracking of rigid head motion during MRI using an EEG system". In: *Magn. Reson. Med.* (2022).
- [28] V. Lohner et al. "Incidental findings on 3 T neuroimaging: cross-sectional observations from the population-based Rhineland Study". In: *Neuroradiology* 64.3 (2022), pp. 503–512.
- [29] M. Jenkinson et al. "Measuring transformation error by RMS deviation". In: *Technical report* (1999).
- [30] H. Peng et al. "Accurate brain age prediction with lightweight deep neural networks". In: *Medical image analysis* 68 (2021), p. 101871.
- [31] G. Huang et al. "Densely connected convolutional networks". In: *CVPR*. 2017, pp. 4700–4708.
- [32] L. Henschel et al. "FastSurfer". In: *Bildverarbeitung für die Medizin 2020*. Springer, 2020, pp. 208–208.

- [33] O. Esteban et al. “MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites”. In: *PloS one* 12.9 (2017), e0184661.
- [34] B. Mortamet et al. “Automatic quality assessment in structural brain magnetic resonance imaging”. In: *Magn. Reson. Med.* 62.2 (2009), pp. 365–372.
- [35] B. Fischl. “FreeSurfer”. In: *Neuroimage* 62.2 (2012), pp. 774–781.
- [36] F. Ségonne et al. “A hybrid approach to the skull stripping problem in MRI”. In: *Neuroimage* 22.3 (2004), pp. 1060–1075.
- [37] C. R. Madan. “Age differences in head motion and estimates of cortical morphology”. In: *PeerJ* 6 (2018).

Appendix C.

Publication: "FastSurfer-LIT: Lesion inpainting tool for whole-brain MRI segmentation with tumors, cavities, and abnormalities"

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. ©2025 The Authors.

FastSurfer-LIT: Lesion inpainting tool for whole-brain MRI segmentation with tumors, cavities, and abnormalities.

Authors: **Clemens Pollak**, David Kügler, Tobias Bauer, Theodor Rüber, and Martin Reuter.

Published in: Imaging Neuroscience, Volume 3 (2025).

DOI: [10.1162/imag_a_00446](https://doi.org/10.1162/imag_a_00446)

Github: [Deep-MI/LIT](https://github.com/Deep-MI/LIT)



FastSurfer-LIT: Lesion inpainting tool for whole-brain MRI segmentation with tumors, cavities, and abnormalities

Clemens Pollak^a, David Kügler^a, Tobias Bauer^{b,c}, Theodor Rüber^{b,c,d}, Martin Reuter^{a,e,f}

^aAI in Medical Imaging, German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

^bDepartment of Neuroradiology, Bonn University Hospital, Bonn, Germany

^cDepartment of Epileptology, Bonn University Hospital, Bonn, Germany

^dCenter for Medical Data Usability and Translation, University of Bonn, Bonn, Germany

^eA.A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston, MA, United States

^fDepartment of Radiology, Harvard Medical School, Boston, MA, United States

Corresponding Author: Martin Reuter (martin.reuter@dzne.de)

ABSTRACT

Resection cavities, tumors, and other lesions can fundamentally alter brain structure and present as abnormalities in brain MRI. Specifically, quantifying subtle neuroanatomical changes in other, not directly affected regions of the brain is essential to assess the impact of tumors, surgery, chemo/radiotherapy, or drug treatments. However, only a limited number of solutions address this important task, while many standard analysis pipelines simply do not support abnormal brain images at all. In this paper, we present a method to perform sensitive neuroanatomical analysis of healthy brain regions in the presence of large lesions and cavities. Our approach called “FastSurfer Lesion Inpainting Tool” (FastSurfer-LIT) leverages the recently emerged Denoising Diffusion Probabilistic Models (DDPM) to fill lesion areas with healthy tissue that matches and extends the surrounding tissue. This enables subsequent processing with established MRI analysis methods such as the calculation of adjusted volume and surface measurements using FastSurfer or FreeSurfer. FastSurfer-LIT significantly outperforms previously proposed solutions on a large dataset of simulated brain tumors ($N = 100$) and synthetic multiple sclerosis lesions ($N = 39$) with improved Dice and Hausdorff measures, and also on a highly heterogeneous dataset with lesions and cavities in a manual assessment ($N = 100$). Finally, we demonstrate increased reliability to reproduce pre-operative cortical thickness estimates from corresponding post-operative temporo-mesial resection surgery MRIs. The method is publicly available at <https://github.com/Deep-MI/LIT> and will be integrated into the FastSurfer toolbox.

Keywords: segmentation, lesion, tumor, software, inpainting, brain filling

1. INTRODUCTION

Neuromorphometry is a ubiquitous method for the analysis of brain MRI, used, for example, for the analysis of longitudinal changes during healthy aging or for group comparisons in clinical trials. While neuroimaging pipelines such as *FreeSurfer* (Fischl, 2012), *FSL* (Jenkinson et al., 2012), or *SPM* (Ashburner, 2009) can produce

results for images with lesions (Radwan et al., 2021) (especially small ones), none of them are developed for, or validated on images with large lesions. In traditional atlas registration-based tools, such as *FreeSurfer* (Fischl, 2012), large errors in brain segmentation or premature termination can occur when abnormal changes in the brain tissue (lesions) make a registration with standardized templates challenging (Radwan et al., 2021). These

Received: 22 June 2024 Revision: 31 October 2024 Accepted: 7 December 2024 Available Online: 7 January 2025



The MIT Press

© 2025 The Authors. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

Imaging Neuroscience, Volume 3, 2025
https://doi.org/10.1162/imag_a_00446

failure modes persist in modern deep learning-based segmentation methods, which tend to perform poorly on data that have not been seen during training (out-of-distribution data). Even if the full brain segmentation is successful, the reconstruction of cortical surfaces is likely compromised when regions of the cortex are damaged, for example, due to invasive tumors or brain resection surgery (see Fig. 1). Meanwhile, the study of disease effects (e.g., glioblastoma effects on overall brain health) and the study of intervention (side-) effects (e.g., impact of radiotherapy, chemotherapy, or surgery) require morphometric analyses on brain MRIs with pathologies. Besides directly studying lesion effects, the blanket exclusion of images with abnormal structure can cause selection bias and decreased statistical power in downstream analysis of association studies. Even when the standard toolboxes complete without failure, the measurements on images containing lesions were found to be biased (Guo et al., 2019). To address this gap, researchers require specialized tools, which can generate accurate morphometric measurements in the presence of such pathologies and cavities.

In previous work, this challenge has been broadly addressed by replacing lesions with healthy looking tissue prior to processing. The process is commonly referred to as inpainting. Specifically, *SynthSR* (Iglesias et al., 2023) and *Virtual Brain Grafting* (VBG) (Radwan et al., 2021) have recently been proposed for inpainting healthy looking tissue in lesion areas. These significant advances enable previously challenging analyses, for example, of “personalized structural connectomics for moderate to severe traumatic brain injury” (Imms et al., 2023), “contralateral alterations in cortical morphology in patients with diffuse low-grade glioma” (Zhang et al., 2022), “structural plasticity of the contralesional hippocampus” (Liu et al., 2023), the role of the hippocampus in

“recovery in persons with post stroke aphasia” (Schevenels et al., 2022), and “tracking the corticospinal tract in patients with high-grade glioma” (Zhylyka et al., 2021). While VBG and SynthSR have unlocked volume and surface-based neuromorphometric analyses for these patient groups, we find that their application can sometimes be unstable, resulting in long runtimes, faulty segmentations, or unreliable estimates of cortical thickness. Additionally, previously proposed methods can only be used on images with a standard resolution of 1 mm, which requires lossy down-sampling of submillimeter MR images and results in decreased fidelity of segmentation, reconstruction, and subsequent analysis (Henschel et al., 2022). Other approaches, such as adding training cases with lesions to deep learning segmentation networks (Weiss et al., 2021), are currently limited, since no datasets with accurate whole-brain segmentations for patients with lesions exist. To cover the whole range of possible brain lesions, a very large and diverse dataset would be required, and even then, a newly trained network and fitting datasets would be required for every application. Robust inpainting, on the other hand, offers a general approach as it provides a synthetically corrected MRI that can be combined with various neuroimaging tools (e.g., also for surface reconstruction or registration purposes). For the *FastSurfer* (Faber et al., 2022; Henschel et al., 2020, 2022) and *FreeSurfer* (Fischl, 2012) toolboxes, for example, accurate inpainting enables both (i) whole-brain segmentation and (ii) cortical surface reconstruction, despite different underlying segmentation and surface reconstruction algorithms.

In this work, we propose *FastSurfer-LIT* a pipeline that performs whole-brain segmentation and surface reconstruction in the presence of small and large lesions and cavities on various resolutions, scanners, and types of lesion (see Fig. 1). The initial step of the pipeline is the

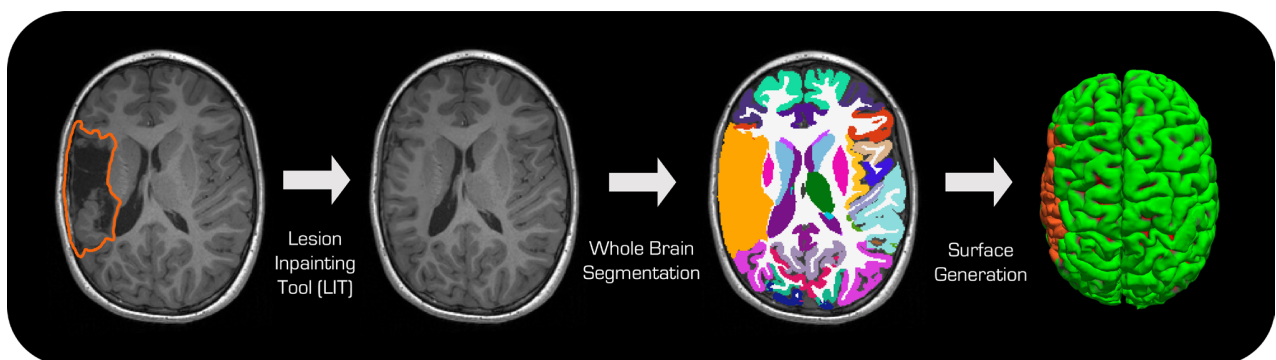


Fig. 1. Overview of FastSurfer-LIT—Based on the two inputs (T1w-image and lesion mask, left), the FastSurfer-LIT pipeline synthesizes a realistic, lesion-free image (center left) by inpainting, then segments the brain into 79 structures (center right) and finally, reconstructs the pial and white matter surfaces (right). The lesion is in orange in segmentations and projected to the surface.

lesion inpainting (LIT) network. To enable high-quality inpainting for arbitrary lesion shapes at multiple desired resolutions, we propose a novel resolution-independent Denoising Diffusion Probabilistic Model (DDPM). This network architecture enables inpainting for any shape or size of mask and can operate on multiple resolutions with high accuracy. By transferring spatial information through the reverse diffusion process, we generate high-quality plausible image content for masked areas, even when substantial parts of the brain are damaged. We combine LIT (the lesion inpainting tool) with FastSurfer (for whole-brain segmentation and surface reconstruction) into the FastSurfer-LIT neuroimaging pipeline and furthermore extend it with new post-processing tools to exclude abnormal regions from volumetric and surface statistical analysis to, for example, enable seamless group analysis including participants with brain lesions.

The evaluation of whole-brain segmentation in the presence of abnormalities is challenging, since no publicly available dataset with (manual) reference segmentations exists. Therefore, one of our experiments employs a tumor growth simulation (Subramanian et al., 2019) in combination with segmentations of lesion-free brains to generate a reference standard for our method. We deform the reference segmentations together with the image according to the tumor growth model to establish a dataset of 158 cases containing synthetic glioblastoma. Additionally, we also generate a second synthetic dataset of multiple sclerosis (MS) lesions ($N = 39$), by transferring lesions masks from patient cases to lesion-free MRI, which again provides us with pairs of reference segmentations and matching lesion masks. Besides these experiments on synthetic data, a blinded rater compared whole-brain segmentations of our method with VBG, the most promising competing method. This evaluation includes 100 cases of hospital patients with 14 different kinds of tumors and surgical cavities. Finally, we use FastSurfer-LIT to compare the consistency of cortical thickness estimates across 14 patients pre- and post-temporo-mesial resection surgery.

In summary, we contribute FastSurfer-LIT, a lesion inpainting, whole-brain segmentation and surface reconstruction pipeline, that

- processes brain MRI containing surgical cavities, tumors, and other lesions independent of their appearance, shape, or size when provided with a corresponding mask,
- performs inpainting natively on high-resolution images,
- outperforms state-of-the-art methods in whole-brain segmentations on MRI with synthetic lesions (and synthetic ground truth), and on MRI from hospital patients (using manual ratings),

- produces cortical thickness estimates with higher consistency across surgical intervention, compared with the state-of-the-art, and
- permits group comparisons to include cases with abnormalities (partially missing data).

These advancements are enabled by a resolution-independent DDPM with a novel slicing scheme for inference, which creates high-quality healthy looking brain MRI individualized for each case.

2. RELATED WORK

2.1. Lesion inpainting methods

Instead of modifying every neuroimaging tool individually to make it robust—considering different lesion types, characteristics, and appearances—previous work has shown that replacing lesions with healthy looking tissue in the input image is an effective and widely applicable strategy. A straightforward approach for inpainting is to replace lesions with intensity values from neighboring areas (Battaglini et al., 2012; Chard et al., 2010; Griffanti et al., 2016; Guo et al., 2019; Magon et al., 2014; Popescu et al., 2014; Prados et al., 2016; Schmidt et al., 2019). Relying on the surroundings of a lesion works well to inpaint smaller areas, but for larger lesions, this approach falls short as whole structures and anatomical detail (e.g., cortical folds) are not recovered. Therefore, methods that work on larger lesions have to rely on a brain model to fill large areas with plausible brain structures. These models can either be learned or explicitly given by a healthy reference brain.

Virtual Brain Grafting (VBG) (Radwan et al., 2021) is a method that transfers healthy looking areas from a reference “donor brain” (Radwan et al., 2021) to a lesion affected area. First, the lesion area is specified by the user, then VBG’s 33-step pipeline transfers matching tissue from the healthy reference to the lesion area and uses FreeSurfer or FastSurfer for segmentation and surface reconstruction. Finally, segmentation maps containing the lesion and other lesion-specific outputs are generated by the pipeline. The initial inpainting process contains skull stripping, non-linear registration, and matching of noise and sharpness levels between the healthy reference and the lesioned brain. In conjunction, these steps often result in a seamless inpainting, which has unlocked the use of FastSurfer and FreeSurfer-based morphometrics for many applications (Imms et al., 2023; Liu et al., 2023; Schevenels et al., 2022; Zhang et al., 2022; Zhylyka et al., 2021). Unfortunately, the registration-based transfer of healthy tissue is itself affected by the appearance of the lesion, which can make the inpainting

less accurate for larger lesions (see, e.g., [Appendix Fig. A3](#)). The tissue from the template brain can also not be adjusted arbitrarily to the target brain, leading to potentially implausible inpaintings for uncommon brain structures. Some of these effects can be mitigated by hand-crafting templates for every population. This, however, is a time consuming process that requires new validation and quality control each time. Since competing methods do not need re-adjustment of templates, we have opted to use the standard template for all cases. Furthermore, in VBG, MRI are always standardized to a voxel resolution of 1 mm, creating interpolation artifacts and losing detail in sub-millimeter MRI, which is becoming widely available ([Glasser et al., 2013](#); [Henschel et al., 2022](#); [Mellerio et al., 2014](#); [Wattjes et al., 2006](#); [Zaretskaya et al., 2018](#)). Finally, the high complexity of the pipeline introduces significant computational costs and adds more than 2.5 hours to the runtime of the FreeSurfer and FastSurfer pipelines, corresponding to a 64% and 192% increase, respectively (see [Section 4.4](#)).

SynthSR ([Iglesias et al., 2023](#)), on the other hand, is a deep learning model, which prepares images for processing with FreeSurfer based on the previously seen training data. The method invokes only a single Convolutional Neural Network (CNN) inference resulting in run times in the order of seconds for inpainting. The training data consist of a combination of T1-weighted scans from the OASIS ([Marcus et al., 2007](#)) dataset as a target and synthetic MRI generated to imitate the appearance of MRI with various resolutions and modalities as input scans. SynthSR standardizes the image, which includes the inpainting of lesions, but also a change of image contrast, orientation, and voxel resolution to 1 mm. This standardization enables use of FreeSurfer with previously unusable clinical images, such as low-resolution, anisotropic voxels, or CT images, as well as some images with lesions. In our experiments we observed that the inpainting of larger lesions is less reliable than smaller lesions (see [Appendix Fig. A3](#)). In contrast to other inpainting methods, SynthSR does not require a lesion mask. While this may seem like a significant advantage, it also removes the guarantee that synthetic intensity changes are limited to the lesion area. Therefore, image changes can occur in unexpected areas (see [Appendix Fig. A2.B,C](#)). It also means that the neural network input includes the lesion itself, making it dependent on the lesion appearance and, thus, more prone to failure for out-of-distribution data. Ultimately, a lesion mask is still required to exclude the synthesized region from downstream analysis.

In addition to SynthSR, various other deep-learning methods for inpainting MR images have been proposed. They are, however, developed for the inpainting of spe-

cific target abnormalities (e.g., removing MS-lesions ([Clèrigues et al., 2023](#); [Tang et al., 2021](#)), removing artifacts ([Xie et al., 2023](#)), re-facing ([Xiao et al., 2022](#)), atlas reconstruction ([Xing et al., 2022](#)), and synthetic validation of images with atrophy ([J. Wang et al., 2023](#))). The aforementioned methods use *Generative Adversarial Networks* (GAN), which were trained by removing areas shaped like the target abnormality from the training images (making them independent of its appearance). The adversarial training paradigm requires a second network to assess the realism of images. A major limitation of GANs is that they are not expected to generalize well to previously unseen mask shapes ([Lugmayr et al., 2022](#)), which means that lesions of unseen size or shape cannot be accurately replaced by these specific methods.

Recently, *Denoising Diffusion Probabilistic Models* (DDPM) have emerged as the state-of-the-art method for natural image inpainting ([Lugmayr et al., 2022](#)). Besides superior performance compared with GANs ([Dhariwal & Nichol, 2021](#); [Lugmayr et al., 2022](#)), these deep learning models can be trained by de-noising images (also called the “reverse diffusion” process), which removes the necessity to generate masks for inpainting during training. During inference, the reverse diffusion process can be leveraged to generate plausible inpainted regions for arbitrary masks, which was previously challenging with GANs ([Lugmayr et al., 2022](#)). They are also independent of the appearance of the inpainted region. DDPMs have now been expanded to 2.5D for 3D inverse problems, such as 3D MRI and CT reconstruction ([Chung et al., 2023](#); [Lee et al., 2023](#)) by applying two 2D diffusion models in two orientations of the volume. This process is similar to view aggregation of 2.5D segmentation models ([Henschel et al., 2020](#); [Roy et al., 2022](#)) and creates coherent volumes during the reverse diffusion process. Because of their strong theoretical advantages and proven utility, we use a DDPM for inpainting in our method. Finally, no resolution-independent DDPM method has been introduced so far creating a gap for sub-millimeter MR acquisition protocols.

2.2. Evaluation of whole-brain segmentation in the presence of large lesions

A general challenge for the development and evaluation of lesion-robust segmentation is the missing ground truth data. While large datasets with manual lesion segmentations exist ([Aerts & Marinazzo, 2018](#); [Aerts et al., 2020](#); [Baid et al., 2021](#); [Bakas et al., 2022](#); [Menze et al., 2015](#)), no database combining lesion segmentations with (manual) whole-brain segmentation has been published to date. Therefore, the authors of VBG evaluated their method based on two datasets:

1. Patient dataset: 10 patients with glioma.
In the absence of ground truth labels, the quality of segmentations of the patient dataset was assessed by experts.
2. Synthetic dataset: 100 synthetic cases from 10 gliomas inserted into 10 “healthy” brain volumes.
The mass effect was mimicked by nonlinear registration of the images. Reference segmentations were generated with FreeSurfer based on deformed images (without lesion insertion).

These two analyses give a good overview of the method performance with reference-based measures such as Dice Similarity Coefficients and expert ratings. However, the patient dataset was limited in size (10) and diversity (only glioma lesions). For the synthetic dataset, the reference standard was generated by FreeSurfer after deformation by the synthetic mass effect, which could lead to decreased quality in the FreeSurfer outputs used as ground truth.

SynthSR was evaluated on three datasets containing lesions:

1. ATLAS (Liew et al., 2018): 655 volumes of stroke patients.
2. BraTS (Menze et al., 2015): 1251 volumes of glioblastoma patients.
3. BraTS-Registration (Baheti et al., 2021): 140 volumes of glioma patients.

While the comparison on these datasets did not include a baseline method, an analysis for the ATLAS dataset of the ipsi- and contralateral volumes of hippocampus, amygdala, thalamus, putamen, and caudate shows asymmetry patterns consistent with the literature. For the lesion inpainting, this was the only analysis of brain segmentations. Additionally, SynthSR was used to aid in the creation of a brain atlas for the BraTS dataset, showing spatial distribution of gliomas consistent with the literature. On the BraTS-Registration dataset, the combination of SynthSR with NiftyReg (Modat et al., 2014) reduced the average landmark error.

3. MATERIALS AND METHODS

3.1. Data

For the training and evaluation of our method, we compile four meta-datasets:

1. No-lesion dataset: To learn the anatomy of brains without large lesions, we combine 11 publicly available datasets into a heterogeneous multi-

resolution dataset of 1750 volumes with isotropic resolutions of 0.7, 0.8, 0.9, and 1 mm. To prevent overoptimistic results by leakage of information from the final test set into the training procedure, we use the same training, validation, and test splits used for the development of FastSurfer (Henschel et al., 2022), resulting in (1315, 80, 355) volumes in the (training, validation, test) set:

- (a) HCP (Glasser et al., 2013) (30, 20, 80)
- (b) RS (Breteler et al., 2014) (30, 20, 80)
- (c) ABIDE-I (Di Martino et al., 2014) (68, 0, 20)
- (d) ABIDE-II (Di Martino et al., 2017) (0, 0, 25)
- (e) ADNI (Jack et al., 2008) (215, 8, 40)
- (f) IXI (“IXI – Information eXtraction from Images”, n.d. (400, 0, 43)
- (g) LA5C (Poldrack et al., 2016) (203, 9, 15)
- (h) MBB (Babayan et al., 2019) (195, 0, 0)
- (i) MIRIAD (Malone et al., 2013) (30, 7, 0)
- (j) OASIS1 (Marcus et al., 2007) (79, 11, 35)
- (k) OASIS2 (Marcus et al., 2010) (65, 5, 17)

To reduce redundancy, details on these datasets can be found in the FastSurferVINN paper (Henschel et al., 2022), where this meta-dataset was first introduced.

2. Synthetic glioblastoma dataset: We use a tumor growth simulation (Subramanian et al., 2019) to generate tumor areas and plausible deformations for 58 randomly selected cases from the healthy validation dataset and 100 from the healthy test set. We detail the simulation process in Section 3.2.
3. Synthetic MS lesion dataset: We use a publicly available dataset of multiple sclerosis (MS) lesions (Commowick et al., 2018, 2021)* and transfer the lesion masks from the original MRI to 39 cases of the healthy test set.
4. Patient dataset: We obtain three datasets with lesion-afflicted images and lesion masks to evaluate our method on clinical cases.
 - (a) UPENN-GBM (Bakas et al., 2021, 2022; Clark et al., 2013): 630 MRI from glioblastoma patients “acquired during routine clinical practice, at the University of Pennsylvania Health System” (Bakas et al., 2022). Images are acquired with multiple scanners, MR sequences, field strengths, and voxel sizes from 0.9 to 5 mm anisotropic. For our analysis, we randomly select 150 isotropic images.

* Data of the 2016 MSSEG challenge dataset were generated by neurologists in OFSEP, the French MS registry (Vukusic et al., 2020). They collect clinical data prospectively in the European Database for MS (EDMUS) software (Confavreux et al., 1992). MRI of patients were provided as part of a care protocol. Nominative data are deleted from MRI before transfer and storage on the Shanoir platform (“SHANOIR SharingNeuroImagingResources”, n.d.).

- (b) BTC (Aerts & Marinazzo, 2018; Aerts et al., 2018, 2020): Overall 44 MRI acquired at the Ghent University Hospital, Belgium. The dataset includes volumes of 11 patients with glioma and 14 patients with meningioma before surgery and 7 and 12 follow-up post-operative scans respectively, all with a voxel size of 1 mm.
 - (c) UKB: 76 MRI of patients acquired at the University Clinic of Bonn, Germany with 11 different types of lesions including hippocampal resections, porencephaly, and Rasmussen's encephalitis acquired at 0.8 mm voxel size.
5. Temporo-mesial resection surgery dataset: 15 pairs of pre- and post-operative scans of patients with mesial temporal lobe epilepsy undergoing temporo-mesial resection surgery acquired at the University Clinic of Bonn, Germany with 0.8 mm voxel size.

Participants of the individual studies gave informed consent in accordance with the Institutional Review Board at each of the participating sites. Complete ethic statements are available at the respective study web pages and cited publications.

3.2. Tumor growth simulation

The lack of manual whole-brain segmentations for images with lesions creates a significant obstacle for quantitative evaluation. To overcome this, we generate MRI with synthetic glioblastoma and accurate segmentation labels by augmenting the validation and test sets of the *no-lesion dataset* with a tumor growth simulation model (Subramanian et al., 2019).

The simulation by Subramanian et al. was shown to create large, realistic deformations mimicking those of real glioblastoma closely (Subramanian et al., 2019). Its inputs are segmentations for gray matter, white matter, cerebrospinal fluid, and ventricles as well as an initial starting point for tumor growth. Required segmentations are generated by FastSurfer based on the no-lesion MRI. The starting point is chosen randomly within the brain mask. For implausible locations, tumor growth is typically minimal and such cases are subsequently excluded. The growth model provides a tumor mask and a deformation warp field mimicking the tumors mass effect. We use the deformations to propagate the reliable whole-brain segmentations created from images without lesions to the images with tumor mask and mass effect. Due to the high computational cost of growth simulation, we limit the synthetic dataset to a subset of (58, 100) of the no-lesion (validation, test) sets. Importantly, we ensure there is no

overlap between data used for method development (i.e., training and validation) and the test of FastSurfer, our LIT network, and the FastSurfer-LIT pipeline.

3.3. Synthetic MS lesion

Similar to synthetic glioblastoma, we also generate a dataset with synthetic MS lesions for the validation of our method. In this case, the mass effect of lesions is negligible, however, the pattern and distribution of lesions in the brain are unique for each individual. Therefore, we map existing lesion masks onto the no-lesion test set via non-linear registration (ANTsPy version 0.5.4 SyN (Avants et al., 2008)). With this strategy, we follow evaluations previously performed for MS-specific inpainting tools (Clèrigues et al., 2023; Tang et al., 2021). We use the resulting 39 cases with synthetic MS lesion masks only for the final evaluation of our method and not during development.

3.4. Diffusion model for inpainting

At the core of our method is the LIT inpainting module which follows the established approach of replacing anomalous areas with healthy looking tissue prior to segmentation and surface reconstruction. Contrary to competing methods (Iglesias et al., 2023; Radwan et al., 2021), we aim to leave regions outside of the marked lesion area completely unmodified. Therefore, we keep image intensities outside of the tumor mask unchanged and also do not alter the image resolution.

We propose a resolution-independent DDPM architecture that can accurately generate images on multiple resolutions (isotropic) via the reverse diffusion process. We base our approach on latent diffusion models (LDM) (Pinaya et al., 2022, 2023; Rombach et al., 2022) competent in generating high-resolution natural images (Rombach et al., 2022) and brain MRI (Pinaya et al., 2023) (for an overview of the general data flow in DDPM inpainting, see Appendix A: Appendix Fig. A1). Contrary to the usual U-Net-like architecture, with only fixed up- and down-sampling by factor 2, we replace one set of up- and down-sampling with VINN (Voxel-size Independent Neural Network) layers. These neural network layers adapt the up- and down-sampling based on the input resolution to standardize the size of feature maps in the latent space. This reduces the voxel-size variance—leading to more effective learning for multi-resolution data.

The vast image size of high-resolution images beyond 256^3 voxels makes the use of fully 3D neural networks impractical (Roy et al., 2022). However, inpainting large abnormalities likely benefits from context information of the whole brain. The intact contralateral hemisphere, for

example, may contain important priors for matching healthy tissue (Xing et al., 2022). In fact, even the background noise may provide important information about the brain’s appearance in the MRI (Pollak, K  gler, Breteler, et al., 2023; Pollak, K  gler, & Reuter, 2023). To use an extended context, while keeping the network and its memory requirements manageable, we combine two strategies: (1) Using three separately trained 2D models for each of the anatomical planes (multi-view) and (2) choosing shifted slabs during each call of the 2D inference with different offsets orthogonal to the current view (varying spatial context). The central idea for both of these strategies is to propagate spatial information via the iterative diffusion process by providing changing information during each network inference. More specifically, our LIT method rotates between using axial, coronal, and sagittal views, similar to Lee et al. (2023). In contrast to Lee et al. (2023), we use slabs of seven neighboring image slices (with full image height and width) as network input, instead of only a single slice. Then we select slabs with different offsets to provide new context for every inference step. While still compatible with the 2D networks, slabs provide additional information in the direction orthogonal to the current view.

Our modifications to the standard DDPM inference scheme and architecture do not require additional computational resources during inference. Furthermore, we only perform inference for slabs containing the lesion area (Lugmayr et al., 2022), which linearly reduces inference cost with lesion size. Overall, the LIT DDPM is an accurate, efficient inpainting method, which natively supports multiple resolutions, uses high spatial context, and is completely independent of the target region’s shape or appearance.

3.5. Whole-brain segmentation and surface reconstruction

For the segmentation and surface reconstruction after inpainting, we employ two popular toolboxes: FastSurfer and FreeSurfer. FastSurfer (version 2.2.0) uses a voxel size independent neural network (VINN) (Henschel et al., 2022) for whole-brain segmentation, while FreeSurfer (Fischl, 2012) (version 7.4.1) is based on a probabilistic atlas segmentation. Both tools provide subsequent white matter and pial surface reconstructions. We further extend FastSurfer and FreeSurfer with functionality to handle lesion areas during statistical analysis of morphometric estimates. Segmentations of the inpainted images are modified retrospectively by replacing them within the inpainted area with a specific lesion mask label. This segmentation is then mapped to the vertices of the pial and white surfaces, by (i) dilating the lesion mask, (ii) marking

the vertices that are intersecting with the surface, and (iii) smoothing the mask border and filling holes on the surface (mode filter). The vertices labeled as lesion mask can then be ignored in downstream statistical group analyses on the participant level (as demonstrated in Section 4.8). This permits statistical analysis even for areas, where lesions are present for some cases in a dataset. While we directly implement LIT into the FastSurfer pipeline, it can additionally be used as a standalone, general inpainting tool and combined with other neuroimaging software.

3.6. Evaluation metrics

We extensively validate FastSurfer-LIT using a variety of metrics for inpainting, segmentation, surface reconstruction, and computational efficiency.

3.6.1. Perceptual similarity

To compare the effect of our modifications to the standard DDPM model, we perform inpainting on the no-lesion dataset using simulated masks and use common perceptual quality metrics to test whether our approach can re-generate the masked areas. First, the Structural Similarity Index Measure (SSIM) (Z. Wang et al., 2004)

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (1)$$

“compares local patterns of pixel intensities that have been normalized for luminance and contrast” (Z. Wang et al., 2004). Here μ_x and μ_y are the mean of x and y , σ_x^2 and σ_y^2 are the variance of x and y , σ_{xy} is the covariance of x and y , finally $c_1 = (0.01 * L)^2$, $c_2 = (0.03 * L)^2$, where L is the dynamic range. To calculate the SSIM, we use PyTorch’s ignite framework (Fomin et al., 2020) with a Gaussian kernel of size 11 and standard deviation of 1.5. The SSIM quantifies the similarity of two image areas between -1 and 1, where 1 is a perfect match, 0 indicates no similarity, and -1 would refer to inverse correlation.

As a second metric, we choose the peak-signal-to-noise ratio (PSNR)

$$PSNR(x, y) = 10 \cdot \log_{10} \left(\frac{M}{MSE(x, y)} \right), \quad (2)$$

where M is the maximum value in the image representation (255 in our case) and MSE is the mean squared error. We also use the ignite framework to calculate this metric. PSNR quantifies the similarity of two image areas x , y in decibel (dB), where higher values indicate a higher similarity.

3.6.2. Segmentation quality

To compare segmentations of the tested method with a reference standard, we use the Dice Similarity Coefficient (DSC) and the Hausdorff Distance (HD). The DSC is defined as

$$DSC(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|}. \quad (3)$$

This metrics show the agreement of two binary masks X and Y . In our case, these masks indicate the location and extent of brain structures. The DSC is zero when there is no overlap between prediction and reference standard. A perfect agreement is indicated by a DSC of 1. The HD is defined as

$$HD(X, Y) = \max \left\{ \sup_{x \in X} d(x, Y), \sup_{y \in Y} d(X, y) \right\}. \quad (4)$$

For the same binary masks X and Y , the HD describes whether the edges of the two structures are close to each other. To get a more robust metric, we do not choose the furthest possible distance, but the 95th percentile. We always indicate the HD in millimeters, where 0 mm means that there is a perfect match of structures up to the 95th percentile. To show whether one method outperforms another significantly, we use the Wilcoxon rank-sum statistic implemented in the SciPy (Virtanen et al., 2020) library. The null hypothesis for this test is that the method ranking is random.

3.6.3. Method failures, runtime, and topological defects

In addition to the reference-based evaluations from previous sections, we analyze three per-run meta-data metrics: success rate, overall runtime, and the surface defect count. Large lesions can cause whole-brain segmentation or surface reconstruction pipelines to crash or fail to produce meaningful outputs. While unsuccessful runs do not affect analysis directly, they can introduce selection biases into downstream analysis. Therefore, we always exclude cases with any failures for any methods (except otherwise stated).

To determine the success rate, we track the runtime of pipelines and terminate instances that do not generate all required outputs within 24 hours. We run this benchmark for a subset of 10 cases of the patient dataset on a dedicated desktop workstation (Intel Xeon W-2245, 64GB RAM, Nvidia Quadro RTX 4000 8GB, solid state drive). We do not use parallel processing or use the machine for other tasks, to avoid interaction effects. Additionally, we

report the overall runtime and runtime of successful cases, since the runtime itself is a valuable criteria due to the associated wait times and energy costs.

Finally, we noticed that faulty or low-quality surface reconstruction is a typical failure mode for all compared pipelines. Such surface errors lead to topological surface defects, pipeline crashes, and long pipeline run times. The number of surface defects is also a common quality measure for quality control of FastSurfer and FreeSurfer runs (Esteban et al., 2017; Rosen et al., 2018). Therefore, we also report the average surface defects on the larger synthetic lesion dataset.

3.6.4. Manual comparison of whole-brain segmentation

In a visual validation, a domain expert compares segmentations derived from the two best-performing inpainting methods: VBG and LIT. This analysis is performed separately when using FreeSurfer or FastSurfer as the segmentation tool. The expert selects the superior segmentation map based on subcortical regions and gray/white matter (GM/WM) boundary for each shown case, while ordering of cases and methods in the viewer is randomized. Additionally, the rater also marks cases as “failure” if multiple major errors occurred, such as missing entire sulci or gyri, mislabeling of the cerebellum as cortex, or completely misplaced corpus callosum.

We randomly draw 50 cases from the patient dataset, choosing from a “random” subset and a “high difference” subset of cases. The high difference cases are selected based on large differences between VBG and LIT-based segmentation maps, which are expected to be more challenging and thus suitable for method comparison (Isensee et al., 2024). More specifically, the cases are selected according to the highest peak value on the Gaussian smoothed segmentation difference map. Since all cases are specific to the used method, these sets differ between FastSurfer and FreeSurfer, and can, therefore, only be used for a direct comparison across inpainting approaches within each segmentation method.

To support the visual inspection during these tasks, we develop a custom rating tool, which guides raters through the process, collects inputs, selects and highlights an area of interest, highlights differences of segmentation maps, etc. (see Appendix C, Appendix Fig. A5). During all rating, a free text field for comments captures the reasoning and supports retrospective analysis of rating decisions. To assess the statistical significance of binary rater decisions (e.g., “Which of two methods is better?”, “Did a method fail?”), we use Fishers exact test (Fisher, 1992), implemented in the SciPy software library (Virtanen et al., 2020).

3.6.5. Consistency of cortical thickness estimates pre- and post-surgery

Finally, we test the consistency of cortical thickness estimates pre- and post-temporo-mesial resection surgery for MRI of patients with epilepsy. Pre-surgery segmentations and surfaces can be generated by the standard neuroimaging pipelines, as existing pipelines are sufficiently robust to temporal lobe atrophy, while post-surgery images require lesion inpainting.

For this test, we smooth all cortical thickness maps (full width at half maximum 15) and map the surfaces onto a common template (fsaverage). Then, we calculate the intraclass correlation coefficient (ICC) (McGraw & Wong, 1996) to determine thickness similarity in the paired samples of pre- and post-surgery images on a vertex level. In this case, ICC indicates the similarity of the thickness values on the surface, where 1 indicates perfect reproducibility. Note, that we use the degree of absolute agreement among measurements (criterion-referenced reliability), which compares equality, not only correlation.

4. RESULTS

In the following section, we first present the effect of our modifications to the standard DDPM inference and architecture. Then, we jointly evaluate our LIT DDPM in two scenarios: (a) inpainting with the FastSurfer pipeline for inpainting (FastSurfer-LIT) and (b) inpainting combined with FreeSurfer for segmentation and surface reconstruction on the same set of experiments. In each scenario, the relevant reference is generated based on lesion-free images by FastSurfer or FreeSurfer, respectively. Additionally, we evaluate multiple method combinations with real-world images, where no reference segmentations are available.

4.1. Method ablation

Initially, we evaluate the effect of our modifications to the baseline 2D DDPM architecture and inference scheme, iteratively removing components of our method and

re-evaluating—until only the baseline method remains (ablation study). While we change the network architecture and inference scheme, the number of network parameters and the required inference steps and input sizes (slabs) are the same across variants. This experiment is performed on the validation split of the synthetic lesion dataset used for method development. First, we run the DDPM-based inpainting in conjunction with the FastSurfer segmentation (Henschel et al., 2022). Then, we calculate SSIM and PSNR to judge the inpainting quality, as well as DSC and HD95 for segmentation performance. Note that the segmentation method is fixed and only the inpainting changes, resulting in different segmentation accuracy. To reduce computational cost during method development, we omit surface reconstruction and generate segmentation maps with the FastSurferVINN neural network only.

The results are shown in Table 1, where the LIT inpainting outperforms other DDPM variants. Since the synthetic lesion only affects the FastSurferVINN inference on volume slices intersecting the lesion, segmentations are identical in most of the volume. In consequence, most structures of the segmentation are not affected and differences of average DSC and HD appear small. Implausible inpainting can, however, have strong impact on the affected slices, and also cause compounding effects further downstream in the processing of FastSurfer and FreeSurfer.

4.2. Synthetic glioblastoma data

We evaluate the accuracy of cortical and sub-cortical segmentations on the synthetic glioblastoma dataset, which contains deformed FreeSurfer and FastSurfer segmentations as a reference standard (see Section 3.2). In the first scenario (using FastSurfer), we compare segmentation performance for (i) no inpainting (baseline) and the two inpainting methods (ii) VBG, and (iii) LIT (ours) (see Fig. 2, left). Both LIT and VBG are developed for compatibility with both FreeSurfer and FastSurfer. We exclude SynthSR, since it has only been evaluated with

Table 1. Method ablation results.

Method configuration	PSNR [dB] ↑	SSIM ↑	DICE ↑	HD95 [mm] ↓
LIT (VINN-DDPM with VA & VSC) (proposed method)	29.64	0.72	0.9492	0.6217
VINN-DDPM with VA	29.30	0.70	0.9491	0.6236
DDPM with VA	29.52	0.72	0.9484	0.6595
Baseline (2D DDPM)	27.91	0.66	0.9482	0.6640

All scores are calculated on the validation set with synthetic lesions and mass effects. Segmentation differences based on FastSurfer appear small, since segmentations perfectly match on slices unaffected by the lesion. Bold values identify the best performing method. LIT is bold to highlight the “final method”.

LIT = Lesion Inpainting Tool (proposed method), VA = view aggregation, VSC = varying spatial context, DDPM = Denoising Diffusion Probabilistic Model, VINN = Voxel Size Independent Neural Network, PSNR = Peak Signal-to-Noise Ratio, SSIM = Structural Similarity Metric, DICE = Dice similarity coefficient, HD95 = 95th percentile Hausdorff distance.

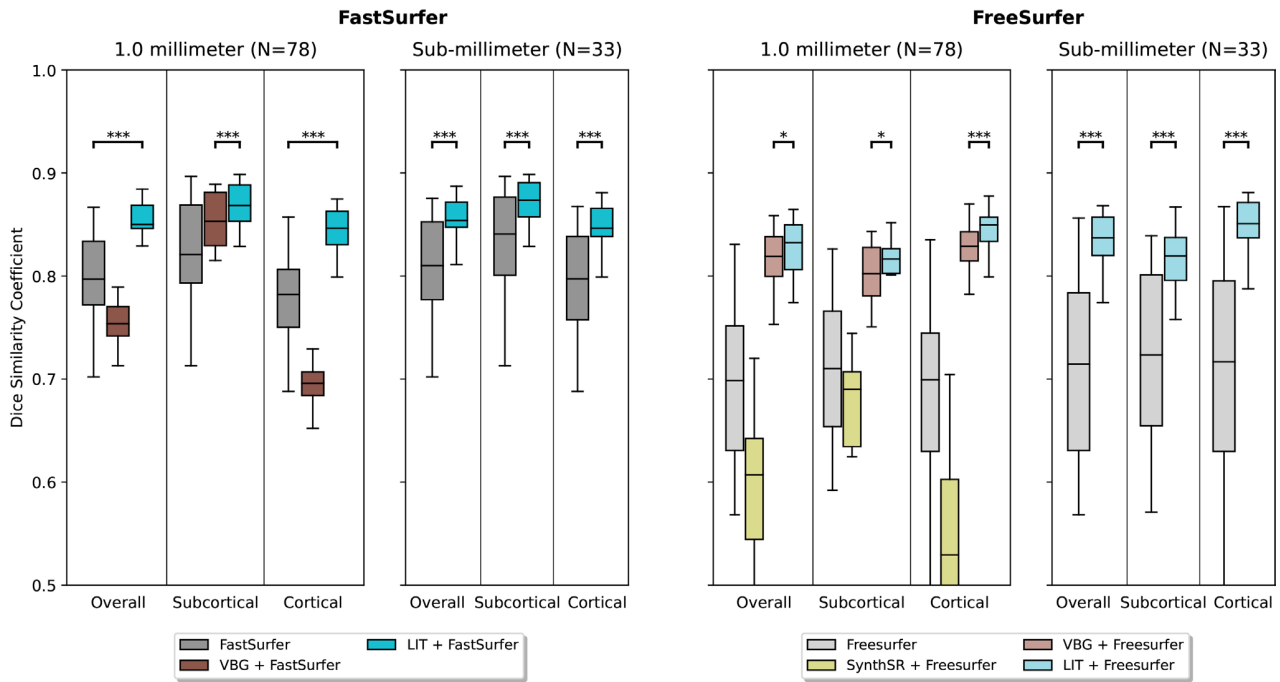


Fig. 2. Method comparison on volumes with simulated glioblastoma. Reference segmentations are generated from lesion-free images by FastSurfer (left) and FreeSurfer (right), respectively (see Section 3.2). Our LIT inpainting significantly outperforms the state-of-the-art on all tests, measured by the Wilcoxon rank-sum statistic. p -Values are indicated for comparison with the second best method. (* $p < 0.05$, ** $p < 0.005$, *** $p < 0.0005$).

FreeSurfer previously and often fails in conjunction with FastSurfer (see Section 4.4). LIT significantly outperforms the state-of-the-art methods on subcortical, cortical, and average DSC. VBG consistently leads to lower scores on cortical DSC, when used together with FastSurfer even performing worse than the unmodified FastSurfer. Mean Hausdorff Distances (HD) of segmentations show the same trends as DSC (no inpainting: 2.16 mm, VBG: 2.69 mm, LIT: 1.34 mm) with all differences of LIT to the second best method significant ($p < 0.0005$). Additionally, we show the segmentation accuracy on volumes with resolution 0.7 mm, 0.8 mm, and 0.9 mm. Here, the only available baseline to compare our FastSurfer-LIT pipeline with is the no inpainting baseline, since the competing methods can only produce segmentations at 1 mm voxel resolution. Our method provides consistent results on sub-millimeter resolution for both DSC and HD (no inpainting: 1.98 mm, LIT: 1.14 mm), indicating robust generalization to high-resolution inpainting and segmentation.

The second scenario illustrated in Figure 2b swaps the FastSurfer pipeline for reference generation and segmentation with FreeSurfer to show the segmentation performance of LIT in a different setting. Here, we compare the effect of (i) no inpainting (baseline), (ii) SynthSR, (iii) VBG, and (iv) LIT (ours) with FreeSurfer segmentations on the same images with synthetic glioblastoma.

Our method outperforms the state-of-the-art methods with statistical significance across the different structures and resolutions. VBG reaches good performance on subcortical DSC and on cortical DSC. SynthSR under-performs FreeSurfer without any modifications on sub-cortical and cortical regions, which might stem from SynthSR modifying and standardizing the contrast of the whole image. The evaluation of Hausdorff Distances (HD) paints a similar picture (no inpainting: 3.27 mm, SynthSR: 4.47 mm, VBG: 1.42 mm, LIT: 1.33 mm). Differences of LIT to the second best method are also statistically significant ($p < 0.0005$). For volumes with sub-millimeter resolution, our method also outperforms the FreeSurfer-only baseline in DSC and HD (no inpainting: 1.93 mm, LIT: 1.13 mm).

4.3. Synthetic multiple sclerosis data

We repeat the analysis of the previous section with a new dataset of MRI with synthetically generated multiple sclerosis lesions. A notable difference between the two datasets is that the synthetic glioblastoma cases include a simulated mass effect, which jointly perturbs images and reference segmentations. This is absent for the synthetic MS lesions, making it a more direct measure of inpainting accuracy, and should yield to higher scores across the board. Additionally, MS lesions are

typically smaller than the simulated glioblastoma, but can be challenging since many lesions can be present in an image and they may be distributed throughout multiple brain regions. In the first scenario (Fig. 3, left), using FastSurfer, with (i) no inpainting, (ii) VBG, and (iii) LIT, our method (LIT) significantly outperforms the two others on 1.0 mm. The same is true for sub-millimeter data, where the (i) no inpainting baseline is the only available comparison. Independent of resolution, our method achieves DSC close to 1 for subcortical structures, indicating supreme inpainting performance. Mean Hausdorff Distances (HD) of segmentations show the same trends as DSC of 1 mm resolution volumes (no inpainting: 0.85 mm, VBG: 2.42 mm, LIT: 0.61 mm) and sub-millimeter volumes (no inpainting: 0.74 mm, LIT: 0.49 mm) with all differences of LIT to the second best method significant ($p < 0.0005$).

For the second scenario, using FreeSurfer, LIT also significantly outperforms the (i) no-inpainting baseline and the competing methods (ii) SynthSR and (iii) VBG, on both 1.0 mm and sub-millimeter MRI. Mean Hausdorff Distances (HD) of segmentations show the same trends for both 1.0 mm resolution (no inpainting: 1.25 mm, VBG: 1.28 mm, SynthSR: 3.02 mm, LIT: 1.14 mm) and sub-millimeter resolution (no inpainting: 1.04 mm, LIT: 0.91 mm). The difference in HD of LIT to the second best method (FreeSurfer) is significant for 1.0 mm resolution

volumes ($p < 0.05$), but not quite for sub-millimeter resolution volumes ($p = 0.057$). We note that inpainting with VBG does not improve segmentation, compared with the no-inpainting baseline, except for cortical regions on 1.0 mm MRI processed with FreeSurfer. VBG has previously only been evaluated on glioma and gliomatous lesions. Our results on synthetic MS lesions indicate that inpainting multiple disconnected regions may be outside of VBGs area of application.

4.4. Processing speed and failures

We evaluate success rate and overall method runtime on a subset of 10 cases of the patient dataset. Here, we test all combinations of the inpainting tools (SynthSR, VBG, LIT) with FastSurfer and FreeSurfer. We show the success rates in Figure 4A. In the first evaluation scenario (using FastSurfer), only the SynthSR variant fails (4/15 cases). We hypothesize that the segmentation failures of SynthSR + FastSurfer stem from an atypical intensity distribution incompatible with FastSurfer (see Appendix Fig. A2). We conclude that the two methods are incompatible and exclude this combination from other evaluations. For the second scenario (using FreeSurfer), failures occur when using vanilla FreeSurfer (no inpainting) (4/15) and SynthSR for inpainting (1/15). For later comparisons (Section 4.6), we additionally run the two most promising

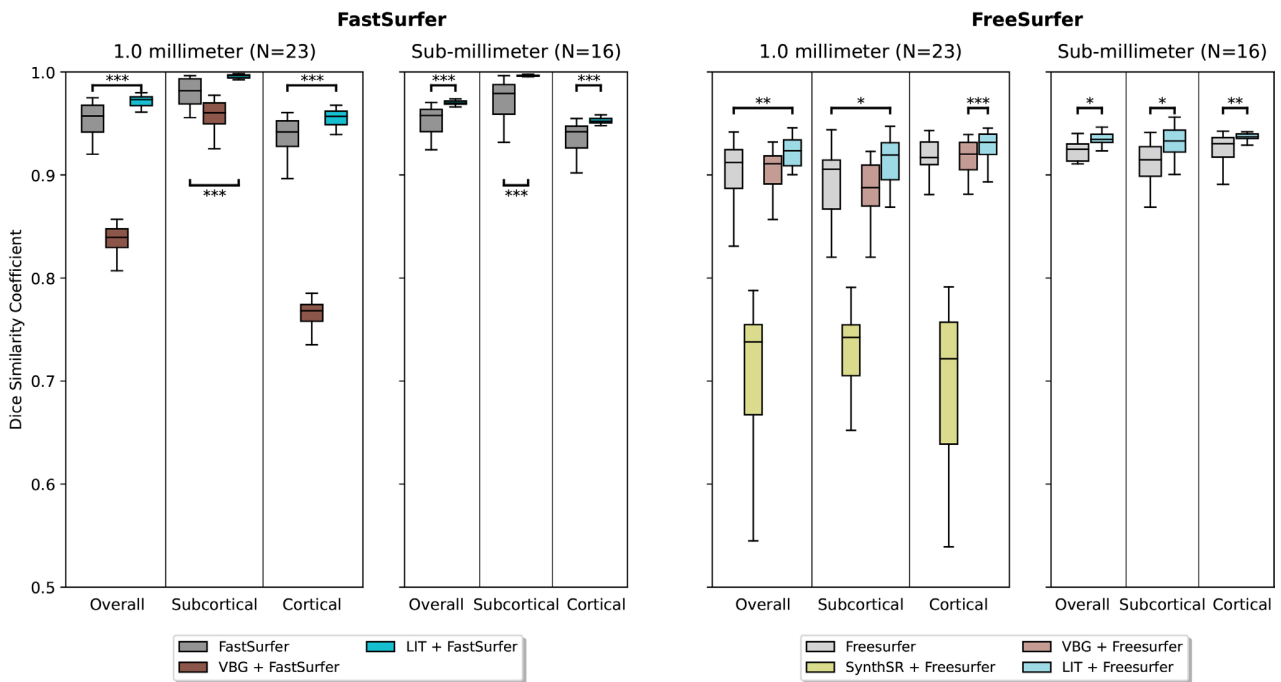


Fig. 3. Method comparison on images with synthetic multiple sclerosis lesions. Reference segmentations are generated from lesion-free images by FastSurfer (left) and FreeSurfer (right), respectively (see Section 3.3). Our LIT inpainting significantly outperforms the state-of-the-art on all tests, measured by the Wilcoxon rank-sum statistic. p -Values are indicated for comparison with the second best method. (* $p < 0.05$, ** $p < 0.005$, *** $p < 0.0005$).

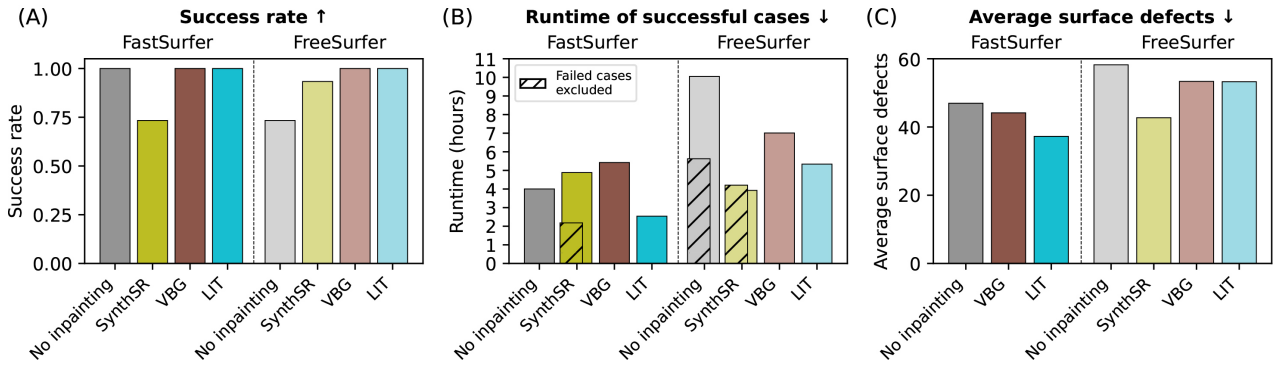


Fig. 4. From left to right: The success rate of method (fraction of segmentations and surfaces generated within 24 hours), runtime on a desktop workstation, number of surface defects before fixing of surface topology. Subfigures A+B are based on 15 cases of the patient dataset, while Subfigure C is based on 100 cases of the synthetic lesion dataset.

methods VBG and LIT on the whole patient dataset, where none of them fails for any of the 270 cases.

We also show the average time to fully process an image including inpainting, segmentation, and surface creation in [Figure 4B](#). For the FastSurfer combinations, LIT is fastest, followed by vanilla FastSurfer (no inpainting), VBG, and SynthSR. The long processing time of SynthSR is driven by the failure cases, but even when looking only at the 11 successful cases, it is still only slightly faster than LIT (average runtime on all 15 cases). VBG has an overall runtime of 5 hours, taking more than twice as long as LIT. Especially for the FastSurfer toolbox, the overhead of the slow template-based VBG inpainting is disproportionate to its fast segmentation and surface reconstruction. For methods run in conjunction with FreeSurfer, SynthSR generates outputs the fastest, followed by LIT, VBG, and vanilla FreeSurfer (no inpainting). In this case, crashes decrease the runtime for SynthSR, as these crashes seem to have happened early in FreeSurfer processing. Vanilla FreeSurfer has a slightly longer runtime on successful cases than LIT (on all cases)—showing that additional inpainting time is largely compensated in LIT by faster downstream processing, for example, during surface creation.

4.5. Surface defects

We assess the quality of surfaces by counting the defects prior to topology fixing in FastSurfer and FreeSurfer on the synthetic glioblastoma dataset and show the results in [Figure 4C](#). For the FastSurfer scenario, LIT inpainting results in the fewest surface defects, followed by VBG and the vanilla FastSurfer baseline (no inpainting). For method combinations with FreeSurfer, SynthSR produces the least surface defects, followed by LIT and VBG with similar number of defects. The FreeSurfer baseline produces the most defects overall. The reduction in sur-

face defects as a result of the inpainting as a preprocessing step causes FreeSurfer and FastSurfer to run faster (shown in [Section 4.4](#))—fewer surface defects require less computationally expensive topology fixing.

4.6. Comparison on patient data

We choose the two previously best performing methods VBG and LIT and let a domain expert compare both methods with FreeSurfer and FastSurfer segmentations using the previously discussed protocol (see [Section 3.6.4](#)).

For the FastSurfer processing, LIT inpainting produces better whole-brain segmentations in 93% of decidable cases (in 18% of all cases no decision was possible, see [Table 2](#)). The same trend holds true for the FreeSurfer segmentations, where LIT is chosen better in 91% of decidable cases (10% no decision). A review of the comments reveals that if no method was decided to be superior, this was most often due to low image quality, high similarity between segmentations, or inaccurate lesion masks.

On the second, failure rating task, FastSurfer in combination with LIT, shows only one failed segmentation for random samples (4%) and no failure (0%) on the high difference set. VBG on the other hand fails 4 times (17%) on random samples and 10 times (37%) on the more challenging high difference cases. For the FreeSurfer-based comparison, LIT fails for no cases (0%) on the random set and for one case on the high difference set (4%), while VBG fails for 4 (15%) and 16 cases (70%), respectively. Overall LIT inpainting is superior combined with both FreeSurfer and FastSurfer methods, providing more accurate segmentations maps in direct comparison and also producing less failures. This is especially true for the high difference set, where low agreement is likely caused by frequent VBG failures and its downstream effects.

4.7. Qualitative results

In [Figure 5](#) we show a representative case, where a cavity was caused by a transsylvian hippocampal resection. On the top left we show the original image (no inpainting). The inpainting of SynthSR fills the lesion area with plausible tissue, however, we can see that the image contrast is changed everywhere, leading to gray matter, that appears thicker than in the original and a general loss of detail in the image. VBG fills parts of the cortical

area with plausible tissue, but produces sharp edges and cavities in the subcortical structures (red arrow). Our LIT inpainting fills the area with plausible tissue and continues structures outside of the mask.

In the resulting FastSurfer segmentations, we observe that FastSurfer without inpainting generates visually plausible results. SynthSRs contrast adjustment, however, leads to generally enlarged gray matter volume for this case. Additionally, a fold with strong over-segmentation was introduced (posterior red arrow). At the front of the gray matter is also over-segmented (anterior red arrow). The FastSurfer segmentation based on VBG inpainting causes gray matter under-segmentation in the posterior regions (e.g., red arrow) and over-segmentation in the anterior region (red arrow). For our LIT inpainting, the posterior regions are segmented as accurate as with the vanilla FastSurfer. In the anterior region, previously challenging for other methods, our method performs as expected, resulting in more accurate segmentation.

For FreeSurfer, the baseline without inpainting contains extremely under-segmented gray matter (posterior red arrow) and erroneous hypointensities. It also contains extreme over-segmentation in the challenging anterior region (red arrow). SynthSR slightly improves the segmentation in the anterior region, but causes the same over-segmentation in the posterior region, as previously seen with FastSurfer. The FreeSurfer segmentation map based on VBG inpainting shows the same issues as the

Table 2. Results of manual quality rating on patient data.

Method	N	VBG+FastSurfer	LIT+FastSurfer
Winning Method	50	3	38
Failures (random)	23	17%	4%
Failures (high diff.)	27	37%	0%

Method	N	VBG+FreeSurfer	LIT+FreeSurfer
Winning Method	50	4	41
Failures (random)	27	15%	0%
Failures (high diff.)	23	70%	4%

In the first row, we compare VBG and our method (LIT) on randomly selected cases and cases selected according to large difference in segmentation maps (high diff.). A blinded rater selected the method with better segmentation maps as “Winning Method.” The rater also annotates segmentation failures shown separately for high diff. cases and others (random). Ratings for VBG and LIT differ significantly in direct method comparison and failures on high difference cases ($p < 0.005$), but not for failures on the random cases.

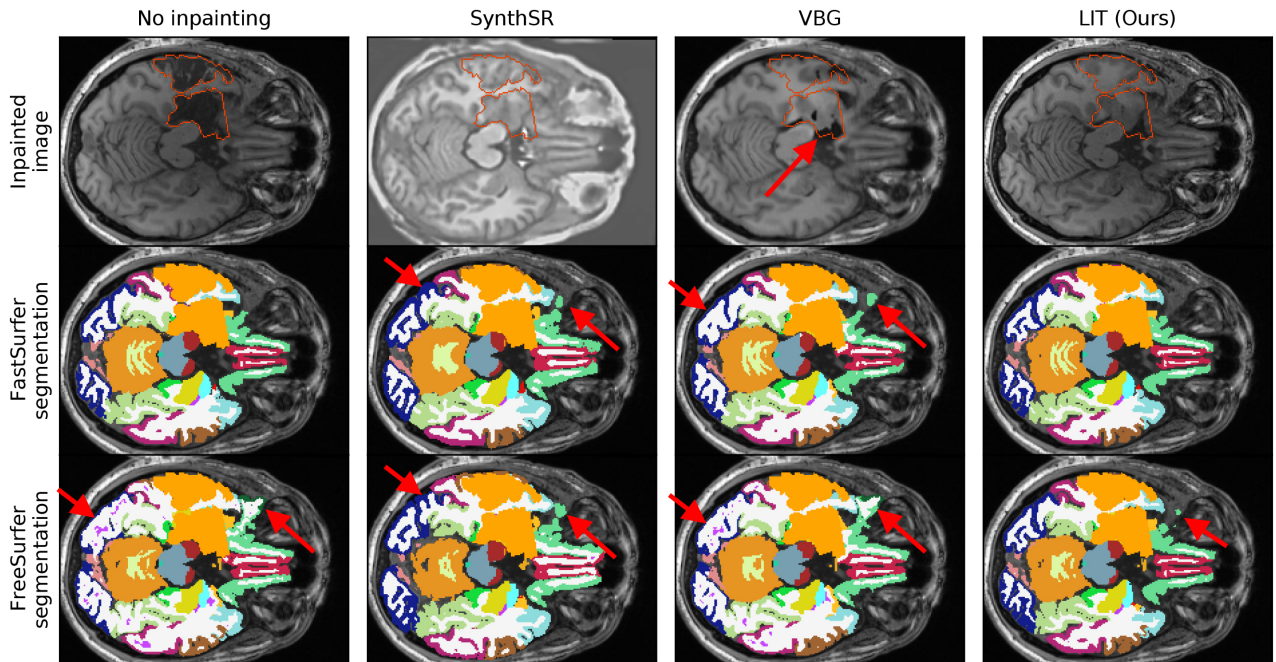


Fig. 5. Qualitative comparison of whole-brain segmentations for all method combinations on a representative case from the UKB dataset. For randomly selected and difficult cases, see [Appendix Figures A2](#) and [A3](#), respectively. The shown slices are located at the center of the lesion. Red arrows indicate inpainting and segmentation flaws.

FreeSurfer segmentation without inpainting. Our LIT inpainting results in accurate segmentation of the posterior regions, and only slight over-segmentation in the anterior region (red arrow).

We show further qualitative results in the [Appendix B: Figures A2](#) (randomly selected cases) and [A3](#) (challenging cases). These cases outline that our method is the only method to modify only the given replacement area ([Appendix Fig. A2.A](#)), accurately inpaints non-brain tissue and skull ([Appendix Figs. A2.D and A3.B](#)), and handles very large abnormalities well ([Appendix Fig. A3.A](#)).

4.8. Cortical thickness estimates before and after surgery

We evaluate the consistency of cortical thickness estimates on MRI from patients before and after undergoing temporomesial resection surgery for VBG and LIT. We show the intraclass correlation coefficients on the surface of a template brain for all methods in [Figure 6](#). As it is reasonable to assume that the cortical thickness does not change between pre- and post-operative scans, the ICC of cortical thickness reflects how robust the cortical thickness analysis is when introducing lesion inpainting. In a best-case scenario, the only difference between cortical thickness estimates are test-retest effects.

For the FastSurfer scenario, we observe generally higher ICC values with LIT than with VBG specifically in frontal and postcentral regions. For the FreeSurfer scenario, we observe the same trend, with the LIT variant also having ICC values close to 1 in several regions (e.g., superior frontal, postcentral, parietal, and occipital). VBG has large areas with very low numbers below 0.3 though. When comparing FastSurfer with FreeSurfer, we see generally higher ICC for FastSurfer, which is consistent with previously reported test-retest results ([Henschel et al., 2020](#)).

5. DISCUSSION

In this work, we introduce FastSurfer-LIT, a whole-brain segmentation and surface reconstruction pipeline for structural MRI with lesions and cavities. As shown in the ablation study, our extension to the latent DDPM architecture and inference improves segmentation quality. We leverage this inpainting quality to improve whole-brain segmentation and surface reconstruction in the presence of lesions. Our pipeline outperforms state-of-the-art methods in four evaluations ranging from experiments on MRI with synthetic lesions, expert ratings on our highly heterogeneous, real-world patient dataset with 14 different types of lesions, and to a consistency of cortical

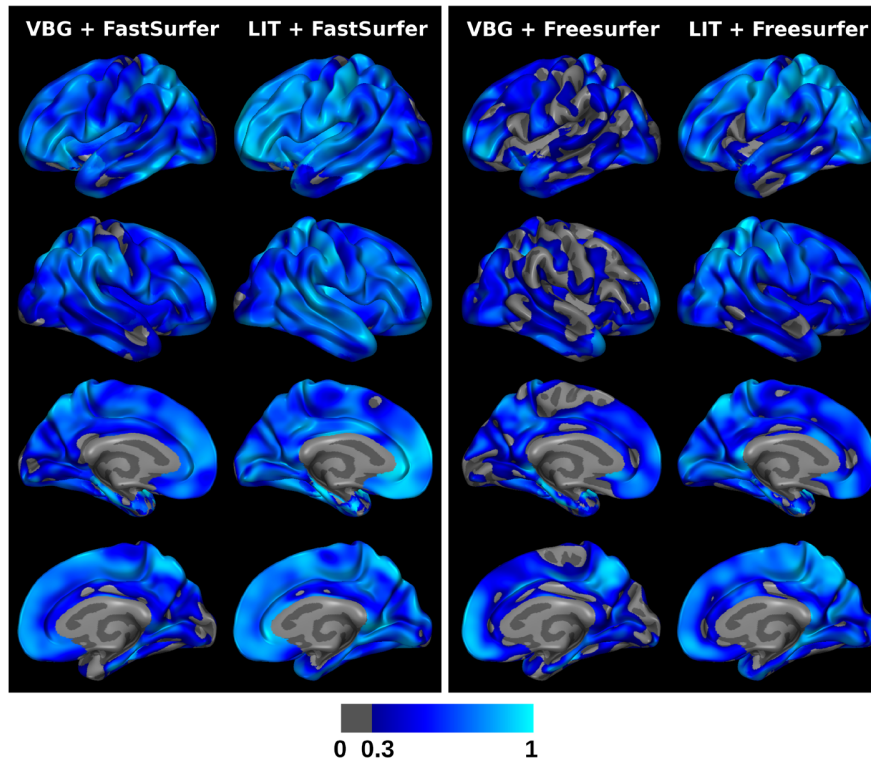


Fig. 6. Comparisons of intraclass correlation coefficient (ICC) values for the cortical thickness of patients before and after temporomesial resection. Each column has different views of the semi-inflated template brain surface, where lighter shades of blue indicate a higher ICC and higher reproducibility of thickness estimates, when using inpainting for lesion filling.

thickness analysis of patient MRI. The variety of lesions and their size in our datasets highlights the robustness and breadth of application of our method (see [Appendix Fig. A3](#) for challenging cases). While competing methods were previously limited to 1 mm voxel resolution, our method can natively inpaint sub-millimeter MRI, which enables analysis for images with higher level of detail. Finally, with 2.5 hours of runtime on average for joint segmentation and cortical reconstruction, or approximately 30 minutes for whole-brain segmentation only, FastSurfer-LIT is an efficient tool for morphometric analysis. While generally much faster than the previous template-based approach (VBG), its speed does not stem from fast inpainting alone (as the iterative reverse diffusion is slower than, for example, the single shot SynthSR), but also from robust and accurate inpainting that enables the segmentation and surface reconstruction to be faster, for example, because of fewer topological surface defects.

The superior performance on challenging cases with either large or unusual lesions is unique to our method, because it is the only one that is designed to be unaffected by the shape or appearance of the lesion. Furthermore, our method guarantees to leave the image outside of the mask completely unchanged during the inpainting process, which is reflected in higher quality segmentations in areas outside of the lesion mask. Finally, our learned brain model allows LIT to generate tissue that is completely individualized to the target brain, leading to more plausible inpaintings than template-based approaches (see [Appendix Fig. A2](#)). A disadvantage of individualized inpainting is that when masks do not cover a whole cavity, the unmasked part might be expanded into the lesion mask to create a plausible inpainting—for such cases we recommend dilating the lesion mask. Fortunately, the expansion of tumors or other distinct pathologies is unlikely, since these cases are not seen during training and are, therefore, not part of the learned brain model.

By integrating FastSurfer-LIT into the FastSurfer project (<https://github.com/Deep-MI/FastSurfer>), we hope to accelerate research on the impact of interventions and disease (e.g., radiotherapy, surgery, glioblastoma) on overall brain health. FastSurfer-LIT could grant new insights into cortical reorganization ([Zhang et al., 2022](#)), with higher resolution data, more accurate cortical reconstruction, and the ability to exclude only the lesion area, instead of limiting the analysis to areas where all patients are lesion free. Besides accelerating and improving research, this work can also contribute to fairness and accessibility of personalized medicine (e.g., personalized structural connectomics ([Imms et al., 2023](#))), which can be made available for individuals with abnormal brain structures.

As segmentation of abnormal areas is out of the scope of this work, masks have to be generated manually or by one of the many available segmentation methods for lesions (as done previously by [Zhang et al. \(2022\)](#)). In the future, FastSurfer-LIT may be extended with more general anomaly detection ([Wolleb et al., 2022](#)) to make the pipeline fully automatic. Alternatively researchers could also generously mask tissue that is not of interest (e.g., one entire hemisphere, similar to [Appendix Fig. A3.A](#)) to make sure lesion tissue is excluded, which removes the requirement for accurate segmentation. Since LIT can run independently from the downstream segmentation or surface reconstruction steps of the pipeline, it can easily be combined with other neuroimaging tools. In our evaluation, we focus on segmentation and surface reconstruction with challenging data. We expect LIT to also enable use of other neuroimage analysis software with lesions, such as structural sub-segmentation ([Estrada et al., 2023](#); [Faber et al., 2022](#)) or subject-to-subject and subject-to-atlas registration (previously demonstrated by SynthSR ([Baheti et al., 2021](#); [Iglesias et al., 2023](#))). Therefore, we release both, a standalone version of the tool (<https://github.com/Deep-MI/LIT/>) and also an integration into the FastSurfer toolbox (<https://github.com/Deep-MI/FastSurfer/>).

Overall, we introduce FastSurfer-LIT, an accurate pipeline for automated neuroimage analysis of brains with cavities, tumors, and other lesions of any size. The pipeline works on multiple resolutions, enabling sub-millimeter analyses for the first time and outperforms previous approaches shown by rigorous analysis.

DATA AND CODE AVAILABILITY

The source code for FastSurfer-LIT will be integrated into the FastSurfer pipeline at <https://github.com/Deep-MI/FastSurfer> and available as standalone inpainting tool at <https://github.com/Deep-MI/LIT>. The rating tool used for the user study is available at https://github.com/Deep-MI/segmentation_labeling.

All MRI datasets of the no-lesion dataset are publicly available and references to the open-source repositories are provided in [Appendix Figure A3](#) of the FastSurferVINN paper ([Henschel et al., 2022](#)). The BTC dataset ([Aerts & Marinazzo, 2018](#); [Aerts et al., 2018, 2020](#)) is available at <https://openneuro.org/datasets/ds001226/versions/00001>, the UPENN-GBM dataset ([Bakas et al., 2021, 2022](#); [Clark et al., 2013](#)) is available at the Cancer Imaging Archive <https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70225642> and the MSSEG dataset ([Commowick et al., 2018, 2021](#)) is available at SHANOIR <https://shanoir.irisa.fr/shanoir-ng/welcome>. The UKB data are not publicly available due to their containing

information that could compromise the privacy of research participants. Data of the Rhineland Study are not publicly available because of data protection regulations. However, access can be provided to scientists in accordance with the Rhineland Study's Data Use and Access Policy. Requests to access the data should be directed to Dr. Monique Breteler at RS-DUAC@dzne.de.

AUTHOR CONTRIBUTIONS

Clemens Pollak: Methodology, Conceptualization, Formal analysis, Investigation, Data Curation, Software, Validation, Writing—original draft, Writing—review & editing, Visualization. David Kügler: Conceptualization, Methodology, Investigation, Supervision, Project administration, Writing—original draft, Writing—review & editing. Tobias Bauer: Data Curation, Resources. Theodor Rüber: Data Curation, Resources, Project administration, Funding acquisition. Martin Reuter: Conceptualization, Methodology, Resources, Writing—original draft, Writing—review & editing, Supervision, Project administration, Funding acquisition.

DECLARATION OF COMPETING INTEREST

The authors do not declare any competing interests.

ACKNOWLEDGEMENTS

This work was supported by DZNE institutional funds, by the Federal Ministry of Education and Research of Germany (031L0206, 01GQ1801), the Helmholtz-AI project DeGen (ZT-I-PF-5-078), by NIH (R01 MH131586, R01 MH130899, R01 AG064027, and P41 EB030006), and by the Chan-Zuckerberg Initiative's Essential Open Source Software for Science RFA (EOSS5 2022-252594). Tobias Bauer received funding from the BONFOR research commission of the medical faculty of the University of Bonn (2022-1A-21) and the Neuro-aCSis Bonn Neuroscience Clinician Scientist Programme (2024-12-07).

We would like to thank the Rhineland Study group (PI Monique M.B. Breteler) for supporting the data acquisition and management. Data used in the preparation of this article were obtained in part by the OASIS Cross-Sectional with principal investigators D. Marcus, R. Buckner, J. Csernansky, J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382, and OASIS: Longitudinal: Principal Investigators: D. Marcus, R. Buckner, J. Csernansky, J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382. Further, data used in the preparation of this article were obtained from the MIRIAD database. The MIRIAD investigators did not participate in

analysis or writing of this report. The MIRIAD dataset is made available through the support of the UK Alzheimer's Society (Grant RF116). The original data collection was funded through an unrestricted educational grant from GlaxoSmithKline (Grant 6GKC). Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. Data collection and sharing for this project were funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. Data were also provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. The acquisition and publication of the UPENN-GBM dataset were partly supported by the National Cancer Institute (NCI), the National Institute of Neurological Disorders and Stroke (NINDS),

and the National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health (NIH) under award numbers NINDS:R01NS042645, NCI:U24CA189523, NCI:U01CA242871, NCATS:UL1TR001878, and by the Institute for Translational Medicine and Therapeutics (ITMAT) of the University of Pennsylvania. The content of this publication is solely the responsibility of the authors and does not represent the official views of the NIH, or the ITMAT of the UPenn. This work was carried out in collaboration with The Observatoire Français de la Sclérose en Plaques (OFSEP), that is supported by a grant provided by the French State and handled by the “Agence Nationale de la Recherche,” within the framework of the “Investments for the Future” program, under the reference ANR-10-COHO-002, by the Eugène Devic EDMUS Foundation against multiple sclerosis and by the ARSEP Foundation.

REFERENCES

- Aerts, H., & Marinazzo, D. (2018). *BTC_preop*. OpenNeuro. <https://doi.org/10.18112/openneuro.ds001226.v5.0.0>
- Aerts, H., Schirner, M., Dholander, T., Jeurissen, B., Achten, E., Van Roost, D., Ritter, P., & Marinazzo, D. (2020). Modeling brain dynamics after tumor resection using the virtual brain. *NeuroImage*, 213, 116738. <https://doi.org/10.1016/j.neuroimage.2020.116738>
- Aerts, H., Schirner, M., Jeurissen, B., Van Roost, D., Achten, E., Ritter, P., & Marinazzo, D. (2018). Modeling brain dynamics in brain tumor patients using the virtual brain. *Eneuro*, 5(3), 1–15. <https://doi.org/10.1523/ENEURO.0083-18.2018>
- Ashburner, J. (2009). Computational anatomy with the SPM software. *Magnetic Resonance Imaging*, 27(8), 1163–1174. <https://doi.org/10.1016/j.mri.2009.01.006>
- Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1), 26–41. <https://doi.org/10.1016/j.media.2007.06.004>
- Babayan, A., Erbey, M., Kumral, D., Reinelt, J. D., Reiter, A. M. F., Röbbig, J., Schaare, H. L., Uhlig, M., Anwender, A., Bazin, P.-L., Horstmann, A., Lampe, L., Nikulin, V. V., Okon-Singer, H., Preusser, S., Pampel, A., Rohr, C. S., Sacher, J., Thöne-Otto, A., ... Villringer, A. (2019). A mind-brain-body dataset of MRI, EEG, cognition, emotion, and peripheral physiology in young and old adults. *Scientific Data*, 6(1), 1–21. <https://doi.org/10.1038/sdata.2018.308>
- Baheti, B., Waldmannstetter, D., Chakrabarty, S., Akbari, M., Bilello, M., Wiestler, B., Schwarting, J., Calabrese, E., Jeffrey, R., Abidi, S., Mousa, M., Villanueva-Meyer, J., Marcus, D. S., Davatzikos, C., Sotiras, A., Menze, B. H., & Bakas, S. (2021). The brain tumor sequence registration challenge: Establishing correspondence between pre-operative and follow-up MRI scans of diffuse glioma patients. *arXiv preprint arXiv:2112.06979*. <https://doi.org/10.48550/arXiv.2112.06979>
- Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F. C., Pati, S., Prevedello, L. M., Rudie, J. D., Sako, C., Shinohara, R. T., Bergquist, T., Chai, R., Eddy, J., Elliott, J., Reade, W., ... Bakas, S. (2021). The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*. <https://arxiv.org/abs/2107.02314>
- Bakas, S., Sako, C., Akbari, H., Bilello, M., Sotiras, A., Shukla, G., Rudie, J. D., Flores Santamaria, N., Fathi Kazerooni, A., Pati, S., Rathore, S., Mamourian, E., Ha, S. M., Parker, W., Doshi, J., Baid, U., Bergman, M., Binder, Z. A., Verma, R., ... Davatzikos, C. (2021). *Multi-parametric magnetic resonance imaging (mpMRI) scans for de novo glioblastoma (GBM) patients from the university of pennsylvania health system (UPENN-GBM) (Version 2)*. The Cancer Imaging Archive. <https://doi.org/10.7937/TCIA.709X-DN49>
- Bakas, S., Sako, C., Akbari, H., Bilello, M., Sotiras, A., Shukla, G., Rudie, J. D., Santamaria, N. F., Kazerooni, A. F., Pati, S., Rathore, S., Mamourian, E., Ha, S. M., Parker, W., Doshi, J., Baid, U., Bergman, M., Binder, Z. A., Verma, R., ... Davatzikos, C. (2022). The university of pennsylvania glioblastoma (UPenn-GBM) cohort: Advanced MRI, clinical, genomics, & radiomics. *Scientific Data*, 9(1), 453. <https://doi.org/10.1038/s41597-022-01560-7>
- Battaglini, M., Jenkinson, M., & De Stefano, N. (2012). Evaluating and reducing the impact of white matter lesions on brain volume measurements. *Human Brain Mapping*, 33(9), 2062–2071. <https://doi.org/10.1002/hbm.21344>
- Breteler, M. M., Stöcker, T., Pracht, E., Brenner, D., & Stirnberg, R. (2014). Ic-p-165: MRI in the rhineland study: A novel protocol for population neuroimaging. *Alzheimer's & Dementia*, 10, P92–P92. <https://doi.org/10.1016/j.jalz.2014.05.172>
- Chard, D. T., Jackson, J. S., Miller, D. H., & Wheeler-Kingshott, C. A. (2010). Reducing the impact of white matter lesions on automated measures of brain gray and white matter volumes. *Journal of Magnetic Resonance Imaging*, 32(1), 223–228. <https://doi.org/10.1002/jmri.22214>
- Chung, H., Ryu, D., McCann, M. T., Klasky, M. L., & Ye, J. C. (2023). Solving 3D inverse problems using pre-trained 2D diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 22542–22551). IEEE. <https://doi.org/10.1109/CVPR52729.2023.02159>
- Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., & Prior, F. (2013). The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26, 1045–1057. <https://doi.org/10.1007/s10278-013-9622-7>
- Clèrigues, A., Valverde, S., Salvi, J., Oliver, A., & Lladó, X. (2023). Minimizing the effect of white matter lesions on deep learning based tissue segmentation for brain volumetry. *Computerized Medical Imaging and Graphics*, 103, 102157. <https://doi.org/10.1016/j.compmedimag.2022.102157>
- Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., Pop, S. C., Girard, P., Améli, R., Ferré, J.-C., Kerbrat, A., Tourdias, T., Cervenansky, F., Glatard, T., Beaumont, J., Doyle, S., Forbes, F., Knight, J., Khademi, A., ... Barillot, C. (2018). Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Scientific Reports*, 8(1), 13650. <https://doi.org/10.1038/s41598-018-31911-7>
- Commowick, O., Kain, M., Casey, R., Ameli, R., Ferré, J.-C., Kerbrat, A., Tourdias, T., Cervenansky, F.,

- Camarasu-Pop, S., Glatard, T., Vukusic, S., Edan, G., Barillot, C., Dojat, M., & Cotton, F. (2021). Multiple sclerosis lesions segmentation from multiple experts: The MICCAI 2016 challenge dataset. *NeuroImage*, 244, 118589. <https://doi.org/10.1016/j.neuroimage.2021.118589>
- Confavreux, C., Compston, D., Hommes, O., McDonald, W., & Thompson, A. (1992). EDMUS, a European database for multiple sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 55(8), 671–676. <https://doi.org/10.1136/jnnp.55.8.671>
- Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34, 8780–8794. <https://doi.org/10.36227/techrxiv.14922633>
- Di Martino, A., O'Connor, D., Chen, B., Alaerts, K., Anderson, J. S., Assaf, M., Balsters, J. H., Baxter, L., Beggato, A., Bernaerts, S., Blanken, L. M. E., Bookheimer, S. Y., Braden, B. B., Byrge, L., Castellanos, F. X., Dapretto, M., Delorme, R., Fair, D. A., Fishman, I., ... Milham, M. P. (2017). Enhancing studies of the connectome in autism using the autism brain imaging data exchange ii. *Scientific Data*, 4(1), 1–15. <https://doi.org/10.1038/sdata.2017.10>
- Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D. A., Gallagher, L., Kennedy, D. P., Keown, C. L., Keyser, C., ... Milham, M. P. (2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 19(6), 659–667. <https://doi.org/10.1038/mp.2013.78>
- Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., & Gorgolewski, K. J. (2017). MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLoS One*, 12(9), e0184661. <https://doi.org/10.1371/journal.pone.0184661>
- Estrada, S., Kügler, D., Bahrami, E., Xu, P., Mousa, D., Breteler, M. M., Aziz, N. A., & Reuter, M. (2023). FastSurfer-HypVINN: Automated sub-segmentation of the hypothalamus and adjacent structures on high-resolution brain MRI. *Imaging Neuroscience*, 1, 1–32. https://doi.org/10.1162/imag_a_00034
- Faber, J., Kügler, D., Bahrami, E., Heinz, L.-S., Timmann, D., Ernst, T. M., Deike-Hofmann, K., Klockgether, T., van de Warrenburg, B., van Gaalen, J., Reetz, K., Romanzetti, S., Oz, G., Joers, J. M., Diedrichsen, J., Giunti, P., Garcia-Moreno, H., Jacobi, H., Jende, J., ... Reuter, M. (2022). Cerebnet: A fast and reliable deep-learning pipeline for detailed cerebellum sub-segmentation. *NeuroImage*, 264, 119703. <https://doi.org/10.1016/j.neuroimage.2022.119703>
- Fischl, B. (2012). FreeSurfer. *NeuroImage*, 62(2), 774–781. <https://doi.org/10.1016/j.neuroimage.2012.01.021>
- Fisher, R. A. (1992). Statistical methods for research workers. In S. Kotz, & N. L. Johnson (Eds.), *Breakthroughs in statistics*. Springer Series in Statistics. Springer, New York, NY. https://doi.org/10.1007/978-1-4612-4380-9_6
- Fomin, V., Anmol, J., Desroziere, S., Kriss, J., & Tejani, A. (2020). High-level library to help with training neural networks in pytorch. <https://github.com/pytorch/ignite>
- Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., & Jenkinson, M. (2013). The minimal preprocessing pipelines for the human connectome project [Mapping the Connectome]. *NeuroImage*, 80, 105–124. <https://doi.org/10.1016/j.neuroimage.2013.04.127>
- Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., Schulz, U. G., Kuker, W., Battaglini, M., Rothwell, P. M., & Jenkinson, M. (2016). Bianca (brain intensity abnormality classification algorithm): A new tool for automated segmentation of white matter hyperintensities. *NeuroImage*, 141, 191–205. <https://doi.org/10.1016/j.neuroimage.2016.07.018>
- Guo, C., Ferreira, D., Fink, K., Westman, E., & Granberg, T. (2019). Repeatability and reproducibility of FreeSurfer, FSL-SIENAX and SPM brain volumetric measurements and the effect of lesion filling in multiple sclerosis. *European Radiology*, 29, 1355–1364. <https://doi.org/10.1007/s00330-018-5710-x>
- Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., & Reuter, M. (2020). FastSurfer—A fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*, 219, 117012. <https://doi.org/10.1016/j.neuroimage.2020.117012>
- Henschel, L., Kügler, D., & Reuter, M. (2022). FastSurferVINN: Building resolution-independence into deep learning segmentation methods—A solution for HighRes brain MRI. *NeuroImage*, 251, 118933. <https://doi.org/10.1016/j.neuroimage.2022.118933>
- Iglesias, J. E., Billot, B., Balbastre, Y., Magdamo, C., Arnold, S. E., Das, S., Edlow, B. L., Alexander, D. C., Golland, P., & Fischl, B. (2023). SynthsR: A public AI tool to turn heterogeneous clinical brain scans into high-resolution T1-weighted images for 3D morphometry. *Science Advances*, 9(5), eadd3607. <https://doi.org/10.1126/sciadv.add3607>
- Imms, P., Clemente, A., Deutscher, E., Radwan, A. M., Akhlaghi, H., Beech, P., Wilson, P. H., Irimia, A., Poudel, G., Dominguez Duque, J. F., & Caeyenberghs, K. (2023). Exploring personalized structural connectomics for moderate to severe traumatic brain injury. *Network Neuroscience*, 7(1), 160–183. https://doi.org/10.1162/netn_a_00277
- Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., & Jaeger, P. F. (2024). nnU-Net revisited: A call for rigorous validation in 3D medical image segmentation. *arXiv preprint arXiv:2404.09556*. https://doi.org/10.1007/978-3-031-72114-4_47
- IXI – Information eXtraction from Images. (n.d.). <https://brain-development.org/ixi-dataset/>
- Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L. Whitwell, J., Ward, C., Dale, A. M., Felmlee, J. P., Gunter, J. L., Hill, D. L., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., ... Weiner, M. W. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27(4), 685–691. <https://doi.org/10.1002/jmri.21049>
- Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., & Smith, S. M. (2012). FSL. *NeuroImage*, 62(2), 782–790. <https://doi.org/10.1016/j.neuroimage.2011.09.015>
- Lee, S., Chung, H., Park, M., Park, J., Ryu, W.-S., & Ye, J. C. (2023). Improving 3D imaging with pre-trained perpendicular 2D diffusion models. *arXiv preprint arXiv:2303.08440*. <https://doi.org/10.1109/ICCV51070.2023.00983>
- Liew, S.-L., Anglin, J. M., Banks, N. W., Sondag, M., Ito, K. L., Kim, H., Chan, J., Ito, J., Jung, C., Khoshab, N., Lefebvre, S., Nakamura, W., Saldana, D., Schmiesing, A., Tran, C., Vo, D., Ard, T., Heydari, P., Kim, B., ... Stroud, A. (2018). A large, open source dataset of stroke anatomical brain images and manual lesion segmentations. *Scientific Data*, 5(1), 180011. <https://doi.org/10.1038/sdata.2018.11>

- Liu, D., Chen, J., Ge, H., Yan, Z., Luo, B., Hu, X., Yang, K., Liu, Y., Xiao, C., Zhang, W., & Liu, H. (2023). Structural plasticity of the contralesional hippocampus and its subfields in patients with glioma. *European Radiology*, 33(9), 6107–6115. <https://doi.org/10.1007/s00330-023-09582-4>
- Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., & Van Gool, L. (2022). RePaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 11461–11471). IEEE. <https://doi.org/10.1109/CVPR52688.2022.01117>
- Magon, S., Gaetano, L., Chakravarty, M. M., Lerch, J. P., Naegel, Y., Stippich, C., Kappos, L., Radue, E.-W., & Sprenger, T. (2014). White matter lesion filling improves the accuracy of cortical thickness measurements in multiple sclerosis patients: A longitudinal study. *BMC Neuroscience*, 15, 1–10. <https://doi.org/10.1186/1471-2202-15-106>
- Malone, I. B., Cash, D., Ridgway, G. R., MacManus, D. G., Ourselin, S., Fox, N. C., & Schott, J. M. (2013). MIRIAD—Public release of a multiple time point Alzheimer's MR imaging dataset. *NeuroImage*, 70, 33–36. <https://doi.org/10.1016/j.neuroimage.2012.12.044>
- Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2010). Open access series of imaging studies: Longitudinal MRI data in nondemented and demented older adults. *Journal of Cognitive Neuroscience*, 22(12), 2677–2684. <https://doi.org/10.1162/jocn.2009.21407>
- Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2007). Open access series of imaging studies (oasis): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 19(9), 1498–1507. <https://doi.org/10.1162/jocn.2007.19.9.1498>
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30. <https://doi.org/10.1037/1082-989X.1.1.30>
- Mellerio, C., Labeyrie, M.-A., Chassoux, F., Roca, P., Alami, O., Plat, M., Naggara, O., Devaux, B., Meder, J.-F., & Oppenheim, C. (2014). 3T MRI improves the detection of transmantle sign in type 2 focal cortical dysplasia. *Epilepsia*, 55(1), 117–122. <https://doi.org/10.1111/epi.12464>
- Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.-A., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., Collins, D. L., Cordier, N., ... Van Leemput, K. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10), 1993–2024. <https://doi.org/10.1109/tmi.2014.2377694>
- Modat, M., Cash, D. M., Daga, P., Winston, G. P., Duncan, J. S., & Ourselin, S. (2014). Global image registration using a symmetric block-matching approach. *Journal of Medical Imaging*, 1(2), 024003. <https://doi.org/10.1117/1.JMI.1.2.024003>
- Pinaya, W. H. L., Graham, M. S., Kerfoot, E., Tudosiu, P.-D., Dafflon, J., Fernandez, V., Sanchez, P., Wolleb, J., da Costa, P. F., Patel, A., Chung, H., Zhao, C., Peng, W., Liu, Z., Mei, X., Lucena, O., Ye, J. C., Tsaftaris, S. A., Dogra, P., ... Cardoso, M. J. (2023). Generative AI for medical imaging: Extending the MONAI framework. *arXiv preprint arXiv:2307.15208*. <https://arxiv.org/abs/2307.15208>
- Pinaya, W. H. L., Tudosiu, P.-D., Dafflon, J., Da Costa, P. F., Fernandez, V., Nachev, P., Ourselin, S., & Cardoso, M. J. (2022). Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models* (pp. 117–126). Springer. https://doi.org/10.1007/978-3-031-18576-2_12
- Poldrack, R., Congdon, E., Triplett, W., Gorgolewski, K., Karlsgodt, K., Mumford, J., Sabb, F., Freimer, N., London, E., Cannon, T., & Bilder, R. (2016). A phenomenon-wide examination of neural and cognitive function. *Scientific Data*, 3(1), 160110. <https://doi.org/10.1038/sdata.2016.110>
- Pollak, C., Kügler, D., Breteler, M. M., & Reuter, M. (2023). Quantifying MR head motion in the rhineland study—A robust method for population cohorts. *NeuroImage*, 275, 120176. <https://doi.org/10.1016/j.neuroimage.2023.120176>
- Pollak, C., Kügler, D., & Reuter, M. (2023). Estimating head motion from MR-images. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)* (pp. 1–5). IEEE. <https://doi.org/10.1109/ISBI53787.2023.10230717>
- Popescu, V., Ran, N., Barkhof, F., Chard, D., Wheeler-Kingshott, C., & Vrenken, H. (2014). Accurate GM atrophy quantification in MS using lesion-filling with co-registered 2D lesion masks. *NeuroImage: Clinical*, 4, 366–373. <https://doi.org/10.1016/j.nicl.2014.01.004>
- Prados, F., Cardoso, M. J., Kanber, B., Ciccirelli, O., Kapoor, R., Wheeler-Kingshott, C. A. G., & Ourselin, S. (2016). A multi-time-point modality-agnostic patch-based method for lesion filling in multiple sclerosis. *NeuroImage*, 139, 376–384. <https://doi.org/10.1016/j.neuroimage.2016.06.053>
- Radwan, A. M., Emsell, L., Blommaert, J., Zhylka, A., Kovacs, S., Theys, T., Sollmann, N., Dupont, P., & Snaert, S. (2021). Virtual brain grafting: Enabling whole brain parcellation in the presence of large lesions. *NeuroImage*, 229, 117731. <https://doi.org/10.1016/j.neuroimage.2021.117731>
- Reuter, M., Tisdall, M. D., Qureshi, A., Buckner, R. L., van der Kouwe, A. J., & Fischl, B. (2015). Head motion during MRI acquisition reduces gray matter volume and thickness estimates. *NeuroImage*, 107, 107–115. <https://doi.org/10.1016/j.neuroimage.2014.12.006>
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10684–10695). IEEE. <https://doi.org/10.1109/CVPR52688.2022.01042>
- Rosen, A. F., Roalf, D. R., Ruparel, K., Blake, J., Seelaus, K., Villa, L. P., Ciric, R., Cook, P. A., Davatzikos, C., Elliott, M. A., Garcia de la Garza, A., Gennatas, E. D., Quarmley, M., Schmitt, J. E., Shinohara, R. T., Tisdall, M. D., Craddock, R. C., Gur, R. E., Gur, R. C., & Satterthwaite, T. D. (2018). Quantitative assessment of structural image quality. *NeuroImage*, 169, 407–418. <https://doi.org/10.1016/j.neuroimage.2017.12.059>
- Roy, S., Kügler, D., & Reuter, M. (2022). Are 2.5D approaches superior to 3D deep networks in whole brain segmentation? Proceedings of the 5th International Conference on Medical Imaging with Deep Learning, *PMLR*, 172, 988–1004. <https://proceedings.mlr.press/v172/roy22a.html>
- Schevenels, K., Michiels, L., Lemmens, R., De Smedt, B., Zink, I., & Vandermosten, M. (2022). The role of the hippocampus in statistical learning and language recovery in persons with post stroke aphasia. *NeuroImage: Clinical*, 36, 103243. <https://doi.org/10.1016/j.nicl.2022.103243>
- Schmidt, P., Pongratz, V., Küster, P., Meier, D., Wuerfel, J., Lukas, C., Bellenberg, B., Zipp, F., Groppa, S., Sämann, P. G., Weber, F., Gaser, C., Franke, T., Bussas,

- M., Kirschke, J., Zimmer, C., Hemmer, B., & Mühlau, M. (2019). Automated segmentation of changes in flair-hyperintense white matter lesions in multiple sclerosis on serial magnetic resonance imaging. *NeuroImage: Clinical*, 23, 101849. <https://doi.org/10.1016/j.nicl.2019.101849>
- SHANOIR SharingNeuroImagingResources. (n.d.). Accessed October 18, 2024. <https://shanoir.org>
- Subramanian, S., Gholami, A., & Biros, G. (2019). Simulation of glioblastoma growth using a 3D multispecies tumor model with mass effect. *Journal of Mathematical Biology*, 79, 941–967. <https://doi.org/10.1007/s00285-019-01383-y>
- Tang, Z., Cabezas, M., Liu, D., Barnett, M., Cai, W., & Wang, C. (2021). LG-Net: Lesion gate network for multiple sclerosis lesion inpainting. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24* (pp. 660–669). Springer. https://doi.org/10.1007/978-3-030-87234-2_62
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17, 261–272. <https://doi.org/10.1038/s41592-019-0686-2>
- Vukusic, S., Casey, R., Rollot, F., Brochet, B., Pelletier, J., Laplaud, D.-A., Sèze, J. D., Cotton, F., Moreau, T., Stankoff, B., Fontaine, B., Guillemin, F., Debouverie, M., & Clanet, M. (2020). Observatoire français de la sclérose en plaques (OFSEP): A unique multimodal nationwide MS registry in France. *Multiple Sclerosis Journal*, 26(1), 118–122. <https://doi.org/10.1177/1352458518815602>
- Wang, J., Larson, K. E., & Oguz, I. (2023). Self-supervised CSF inpainting with synthetic atrophy for improved accuracy validation of cortical surface analyses. *arXiv preprint arXiv:2303.05777*. <https://doi.org/10.3389/fnimg.2022.861687>
- Wang, Z., Bovik, A., Sheikh, H., & Simoncelli, E. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600–612. <https://doi.org/10.1109/TIP.2003.819861>
- Wattjes, M., Harzheim, M., Kuhl, C., Gieseke, J., Schmidt, S., Klotz, L., Klockgether, T., Schild, H., & Lutterbey, G. (2006). Does high-field MR imaging have an influence on the classification of patients with clinically isolated syndromes according to current diagnostic MR imaging criteria for multiple sclerosis? *American Journal of Neuroradiology*, 27(8), 1794–1798. <https://doi.org/10.1007/s00415-007-0666-9>
- Weiss, D. A., Saluja, R., Xie, L., Gee, J. C., Sugrue, L. P., Pradhan, A., Bryan, R. N., Rauschecker, A. M., & Rudie, J. D. (2021). Automated multiclass tissue segmentation of clinical brain MRIs with lesions. *NeuroImage: Clinical*, 31, 102769. <https://doi.org/10.1016/j.nicl.2021.102769>
- Wolleb, J., Bieder, F., Sandkühler, R., & Cattin, P. C. (2022). Diffusion models for medical anomaly detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 35–45). Springer. https://doi.org/10.1007/978-3-031-16452-1_4
- Xiao, Y., Ashbee, W., Calhoun, V. D., & Plis, S. (2022). Refacing defaced MRI with PixelCNN. In *2022 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–7). IEEE. <https://doi.org/10.1109/IJCNN55064.2022.9891937>
- Xie, K., Gao, L., Zhang, H., Zhang, S., Xi, Q., Zhang, F., Sun, J., Lin, T., Sui, J., & Ni, X. (2023). GAN-based metal artifacts region inpainting in brain MRI imaging with reflective registration. *Medical Physics*, 51(3), 2066–2080. <https://doi.org/10.1002/mp.16724>
- Xing, F., Liu, X., Kuo, C.-C. J., El Fakhri, G., & Woo, J. (2022). Brain MR atlas construction using symmetric deep neural inpainting. *IEEE Journal of Biomedical and Health Informatics*, 26(7), 3185–3196. <https://doi.org/10.1109/JBHI.2022.3149754>
- Zaretskaya, N., Fischl, B., Reuter, M., Renvall, V., & Polimeni, J. R. (2018). Advantages of cortical surface reconstruction using submillimeter 7T MEMPRAGE. *NeuroImage*, 165, 11–26. <https://doi.org/10.1016/j.neuroimage.2017.09.060>
- Zhang, S., Sun, H., Yang, X., Wan, X., Tan, Q., Li, S., Shao, H., Su, X., Yue, Q., & Gong, Q. (2022). An MRI study combining virtual brain grafting and surface-based morphometry analysis to investigate contralateral alterations in cortical morphology in patients with diffuse low-grade glioma. *Journal of Magnetic Resonance Imaging*, 58(3), 741–749. <https://doi.org/10.1002/jmri.28562>
- Zhylyka, A., Sollmann, N., Kofler, F., Radwan, A., De Luca, A., Gempt, J., Wiestler, B., Menze, B., Krieg, S. M., Zimmer, C., Kirschke, J. S., Sunaert, S., Leemans, A., & Pluim, J. P. W. (2021). Tracking the corticospinal tract in patients with high-grade glioma: Clinical evaluation of multi-level fiber tracking and comparison to conventional deterministic approaches. *Frontiers in Oncology*, 11, 761169. <https://doi.org/10.3389/fonc.2021.761169>

APPENDIX A

A. INPAINTING DATA FLOW

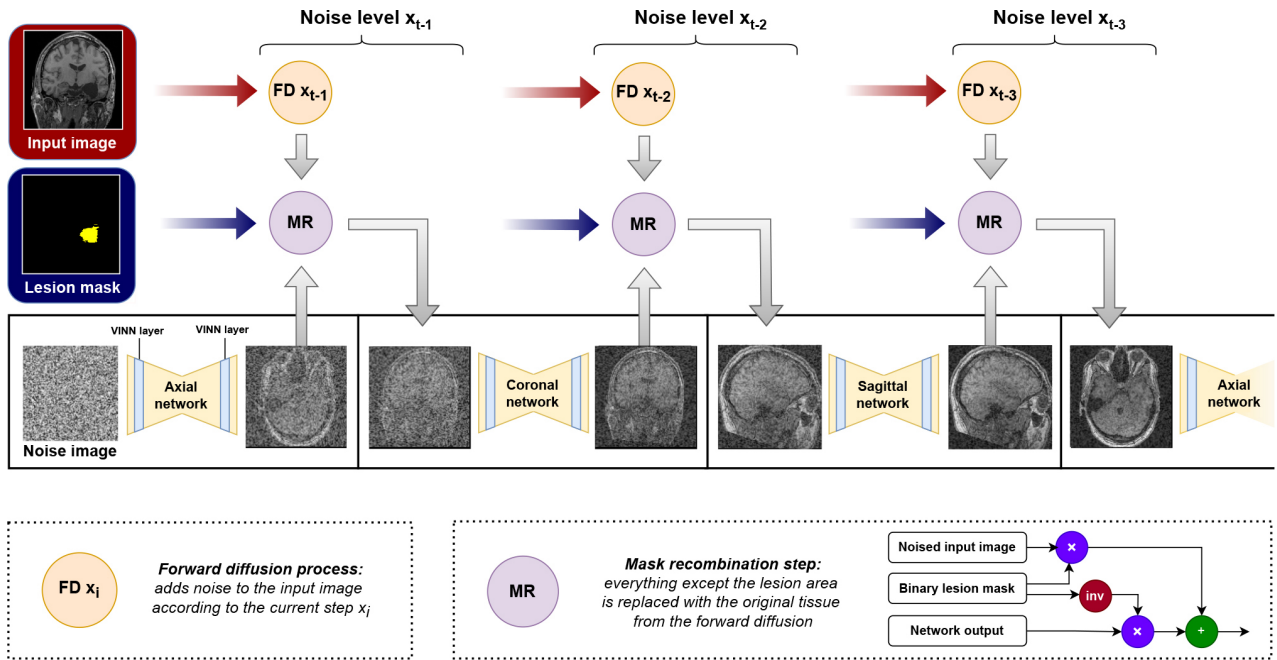
In this work, we train the DDPM U-Net (Pinaya et al., 2022, 2023; Rombach et al., 2022) on lesion-free images. During the inference, we condition the network by continuously replacing the area outside of the lesion mask with a noised portion of the original image. We visualize this process in Appendix Figure A1 and also point to previous publications in the field of computer vision for in-depth background (Dhariwal & Nichol, 2021; Lugmayr et al., 2022). The implementation details for the neural network

and its training can be found in the Open Source repository: <https://github.com/Deep-MI/LIT>.

APPENDIX B

B. QUALITATIVE INPAINTING RESULTS

Some quality metrics for inpainting struggle to capture the complexity of the task. Inpainting regions should seamlessly match the surrounding area (including structural integrity, noise and intensity levels) and also produce plausible brain structures that depend on the



Appendix Fig. A1. Overview of the DDPM inpainting process. The architecture for the axial, coronal, and sagittal networks is shared, but weights differ. Mask recombination is used to guide the inpainting process. Denoising per step is exaggerated in the figure for better viewability. Not shown are slab selection and slab-wise denoising.

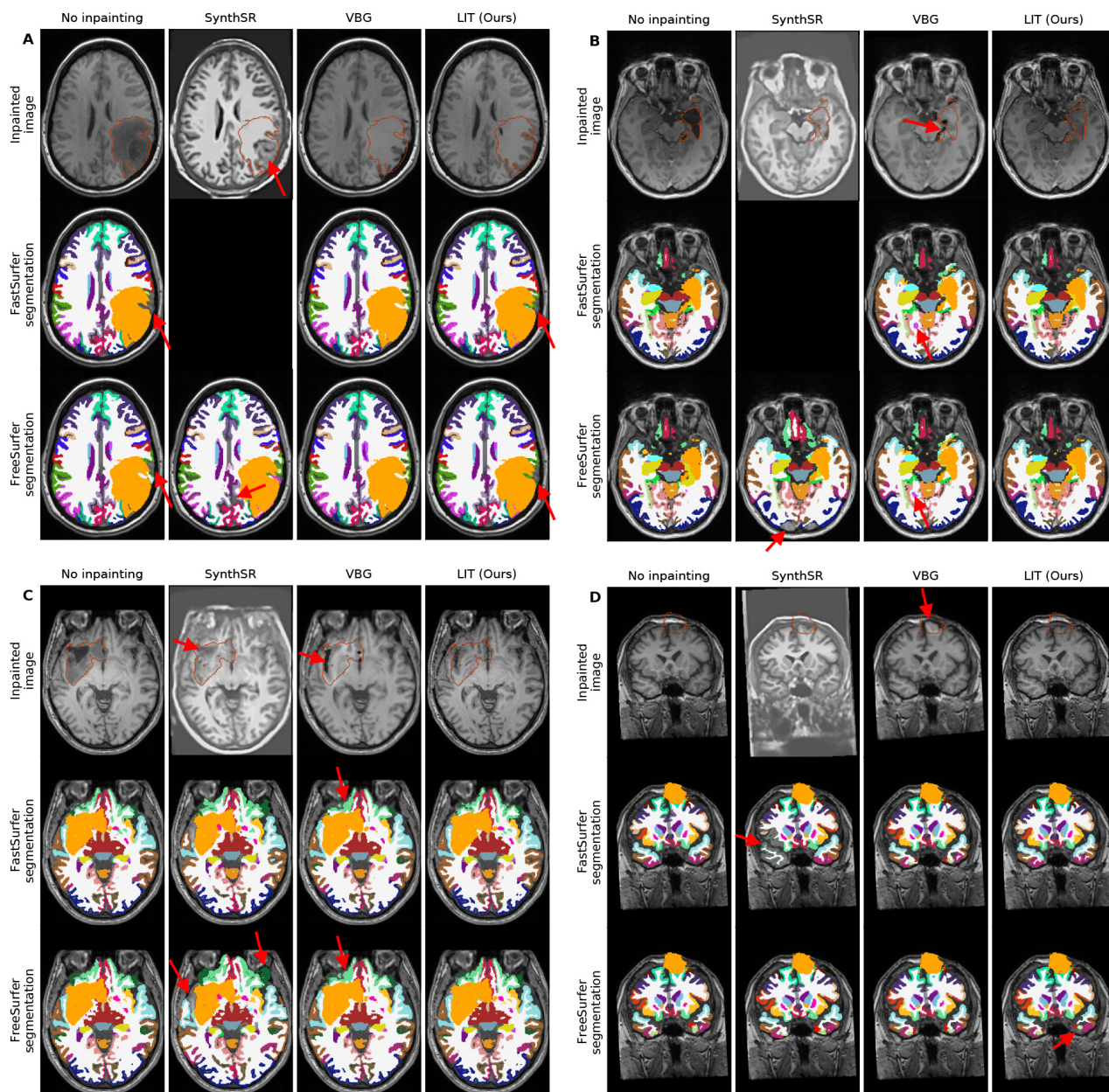
position of the inpainted region, and also on the appearance of the rest of the brain. For a qualitative comparison, we show additional inpainted images with segmentation maps in [Appendix Figures A2–A4](#). First we show a randomly selected case from each of the clinical dataset in [Appendix Figure A2](#). In the inpainted images on the top rows, we observe that the competing methods can modify the image outside of the inpainting mask, which is undesirable and avoided by LIT. SynthSR struggles especially with large lesions (A.2.A) and VBG with complex shapes (A.2.B) and non-brain tissue (A.2.D). In [Appendix Figure A3](#), we show extremely challenging cases, where the competing methods do not produce (meaningful) inpaintings. Meanwhile our method performs as expected. Finally, in [Appendix Figure A4](#), we show that our inpainting network also works in the presence of motion artifacts and low signal-to-noise ratio. While the inpainting can enable use of certain tools, we recommend excluding cases with extreme motion artifacts from analysis with FreeSurfer and FastSurfer, since artifacts are known to bias downstream analysis tasks ([Reuter et al., 2015](#)).

APPENDIX C

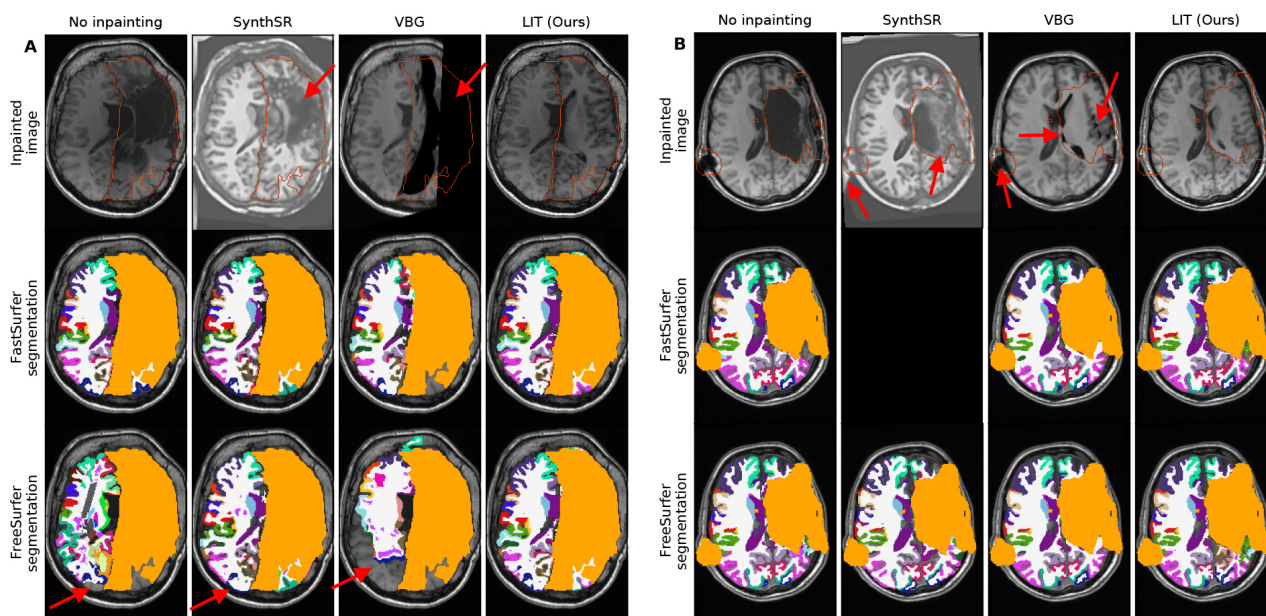
C. RATING PROCEDURE

Judging and comparing the quality of whole-brain segmentation can be a very time consuming process.

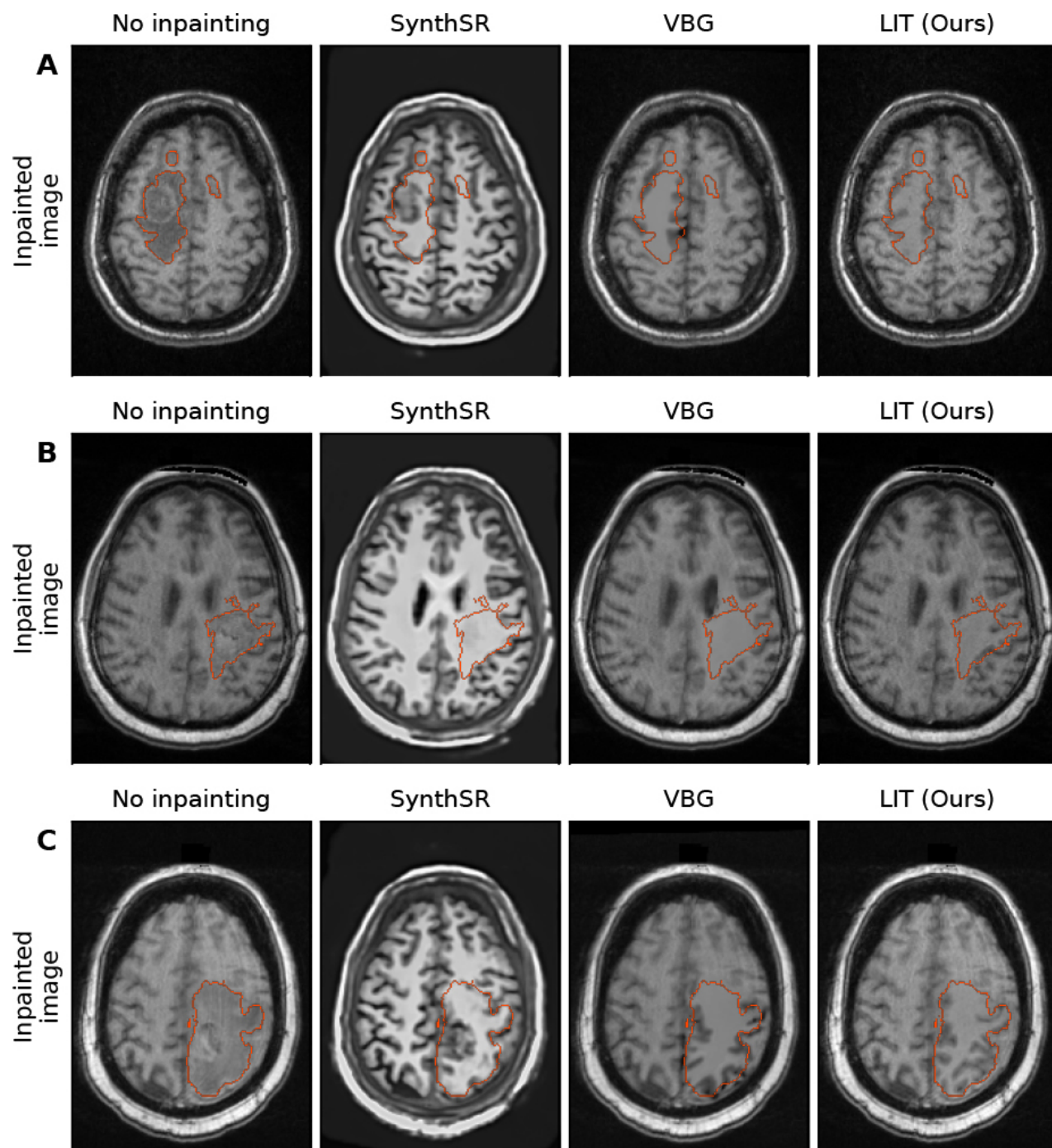
To compare two methods that are giving segmentation of similar quality, raters would typically go through the whole-brain volume and inspect all the different areas. Then they would often identify one or more areas, which show large differences and finally base their judgment on these areas of interest. Based on these observations, we develop an open-source tool and streamline the rating process for the comparison of segmentation maps (shown in [Appendix Fig. A5](#), available on GitHub (https://github.com/Deep-MI/segmentation_labeling)). The goal is to create reliable labels while reducing the required inspection time. This allows us to include more images into this analysis than previous methods ([Radwan et al., 2021](#)). We aim to (i) provide a responsive user interface and (ii) focus the attention of raters. Our rating tool is based on FreeSurfer’s freeview, which is an MR image viewer well known in the community and often does not require additional training of raters. We open and close it from a software package that manages the rating process. To speed up waiting times, all cases are pre-processed and pre-loaded. In addition to the original MRI and the two to-be-compared segmentations maps, we also show a *focus area* and a *difference map*, which can aid the blinded raters by drawing their attention to differences between methods. The *focus area* highlights a suggested area for inspection and decision making. The *difference map* can be switched on to indicate where the segmentation maps disagree, and even spot subtle differences.



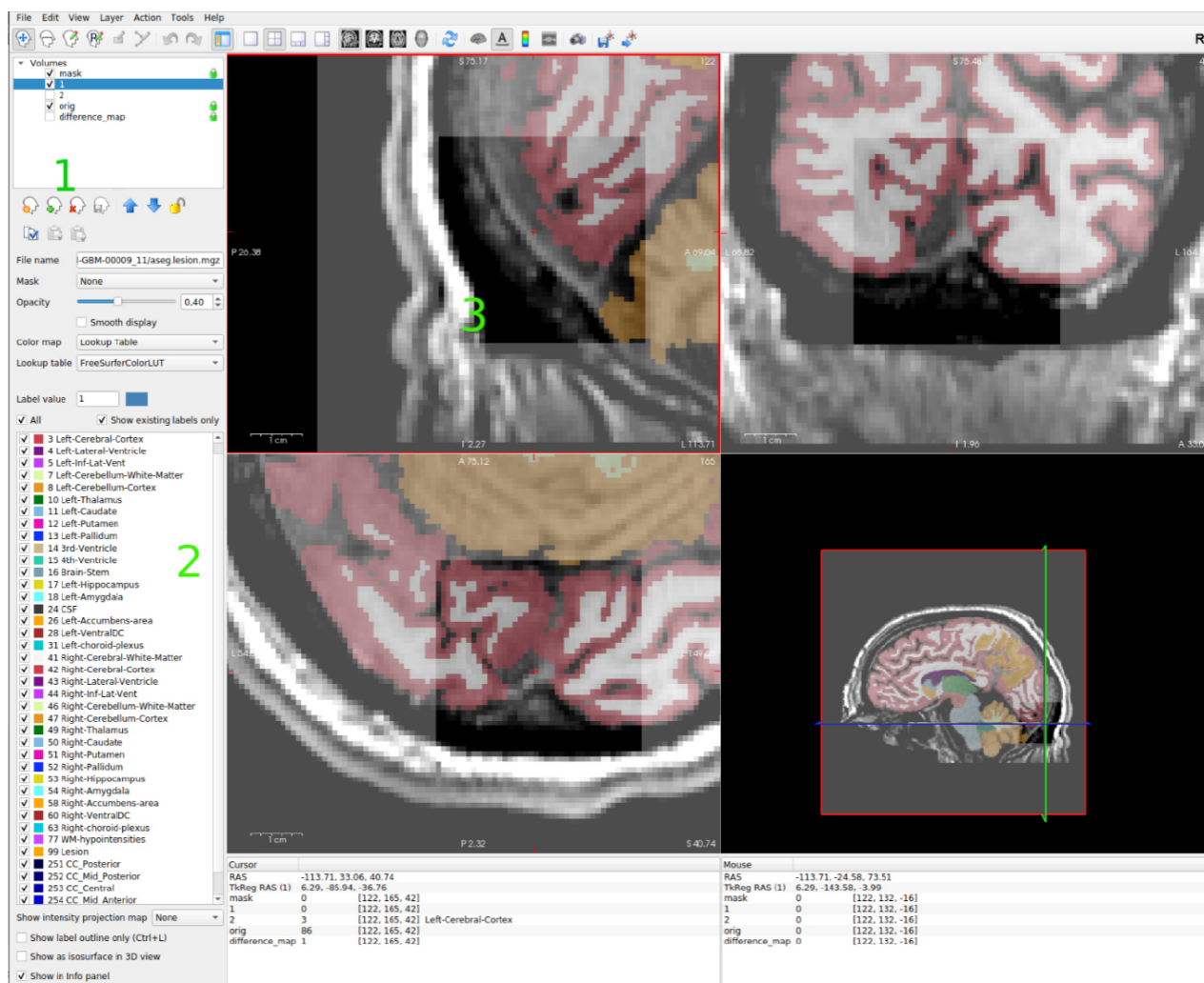
Appendix Fig. A2. Qualitative comparison of whole-brain segmentations for all method combinations on cases randomly selected from the UPENN (A), UKB (B), preoperative BTC (C), and postoperative BTC (D) dataset. Empty spots for SynthSR + FastSurfer represent method failures. The shown slices are located at the center of the inpainting mask. Red arrows point at obvious inpainting and segmentation flaws.



Appendix Fig. A3. Qualitative comparison of two very challenging cases. Shown are whole-brain segmentations for all method combinations. Empty spots for SynthSR + FastSurfer represent method failures. The shown slices are located at the center of the inpainting mask. Red arrows point at obvious inpainting and segmentation flaws.



Appendix Fig. A4. Qualitative comparison of cases with reduced image quality due to motion artifacts. Shown are the original images and inpainted images produced by SynthSR, VBG, and LIT. The shown slices are located at the center of the inpainting mask.



Appendix Fig. A5. Screenshot from the window for the comparison of whole-brain segmentations. The green numbers denote (1) the area in Freeview to manage loading volumes, showing segmentations as 1/2, (2) an overview over some of the segmentation in the label map, (3) the focus area, which is an area of low inter-method agreement pointed out to the rater.

Appendix D.

Publication: "FastSurfer-CC: A robust, accurate and comprehensive framework for corpus callosum morphometry"

This is an author manuscript distributed under the terms of the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. ©2025 The Authors.

FastSurfer-CC: A robust, accurate, and comprehensive framework for corpus callosum morphometry.

Authors: **Clemens Pollak**, Kersten Diers, Santiago Estrada, David Kügler, and Martin Reuter.

Published in: arXiv preprint, under review.

DOI: [arXiv.2511.16471](https://doi.org/10.26434/chemrxiv-2025-16471)

Github: [Deep-MI/LIT](#)

FastSurfer-CC: A robust, accurate, and comprehensive framework for corpus callosum morphometry

Clemens Pollak¹, Kersten Diers¹, Santiago Estrada¹, David Kügler¹, and Martin Reuter^{1,2,3,*}

¹*AI in Medical Imaging, German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany*

²*A.A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston, MA, USA*

³*Department of Radiology, Harvard Medical School, Boston, MA, USA*

*Corresponding author: Martin.Reuter@dzne.de

November 21, 2025

Abstract

The corpus callosum, the largest commissural structure in the human brain, is a central focus in research on aging and neurological diseases. It is also a critical target for interventions such as deep brain stimulation and serves as an important biomarker in clinical trials, including those investigating remyelination therapies. Despite extensive research on corpus callosum segmentation, few publicly available tools provide a comprehensive and automated analysis pipeline. To address this gap, we present FastSurfer-CC, an efficient and fully automated framework for corpus callosum morphometry. FastSurfer-CC automatically identifies mid-sagittal slices, segments the corpus callosum and fornix, localizes the anterior and posterior commissures to standardize head positioning, generates thickness profiles and subdivisions, and extracts eight shape metrics for statistical analysis. We demonstrate that FastSurfer-CC outperforms existing specialized tools across the individual tasks. Moreover, our method reveals statistically significant differences between Huntington’s disease patients and healthy controls that are not detected by the current state-of-the-art.

Keywords: Corpus callosum, Segmentation, Localization, Deep Learning, Head pose, Commissure

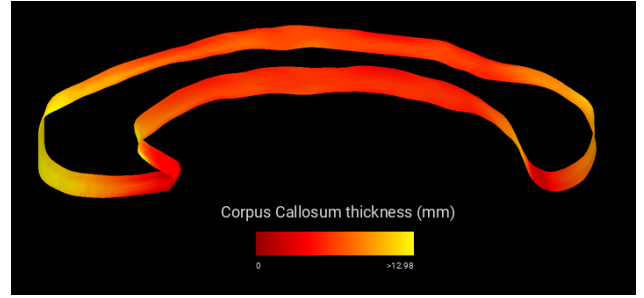
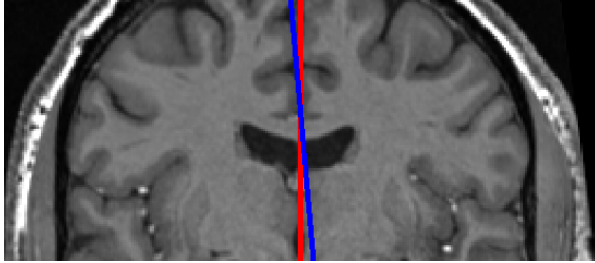


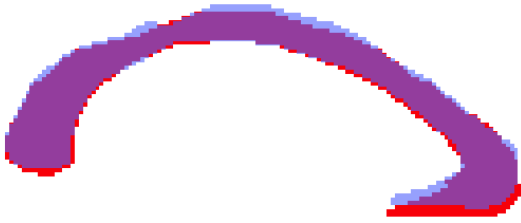
Figure 1: Thickness of an example corpus callosum, shown on a 3D surface model.

1 Introduction

The corpus callosum and fornix, as well as anterior and posterior commissures, are white matter bundles central to the communication between hemispheres, memory recall tasks, and olfaction. In particular, the corpus callosum is the largest brain commissure and is associated with many diseases, for example, epilepsy (Unterberger et al., 2016), autism (Piven et al., 1997), schizophrenia (Woodruff et al., 1995), multiple sclerosis (Ozturk et al., 2010), cerebral palsy (Jaatela et al., 2023), Parkinson’s disease (Yang et al., 2023), bipolar disorder (Grande et al., 2016), and Alzheimer’s disease (Di Paola et al., 2010). The corpus callosum (CC) is an extensively



(a) Two mid-sagittal plane estimates (red & blue) differ only by a rotation of 5° in the coronal plane.



(b) The red mask results from accurate segmentation using the red plane as mid-sagittal plane, while the blue mask results from accurate segmentation using the blue plane.

Figure 2: Choice of mid-sagittal plane affects the shape and thickness of corpus callosum segmentation.

studied structure, which is presented with high contrast in structural MR imaging and shows complex anatomical changes in healthy aging and disease (Luders et al., 2010). Adjacent to the CC lies the body of the fornix (FN), which is part of the limbic system and does not connect the hemispheres directly but forms two C-shaped arches that connect to the hippocampus on the left and right hemispheres and merge below the corpus callosum. The fornix is most prominently involved in Alzheimer’s disease (Lacalle-Aurioles & Iturria-Medina, 2023; Nowrangi & Rosenberg, 2015) and mild cognitive impairment (Nowrangi & Rosenberg, 2015; Zhuang et al., 2012).

The anterior commissure (AC) connects the orbitofrontal, temporal, parietal, and occipital lobes, as well as the insular and entorhinal cortices, with the olfactory bulbs, the septal area, and the amygdalae (Çavdar et al., 2021; Pinto et al., 2017; Raybaud, 2010). The pos-

terior commissure (PC) has been identified as the connection between pre- and postcentral gyri, the superior parietal region in the left hemisphere to the temporal region, and the lateral occipital and superior parietal regions of the contra-lateral hemisphere (Pinto et al., 2017), and is also connected to thalamic nuclei, superior colliculus, and the habenular nuclei at its origins (Pinto et al., 2017). Despite their wide reach, the AC & PC present themselves as distinct white matter tracts almost completely surrounded by gray matter in the mid-sagittal view. Although the function of AC & PC is not well understood, their contrast to the surrounding gray matter has made them essential neuroanatomical landmarks used for localization of neuroanatomical structures during deep brain stimulation (DBS), stereotactic and functional surgery (Baudo et al., 2022), the definition of corpus callosum sub-segments (Jäncke et al., 1997), as well as the standardization of head pose for automated morphometric analysis (C. Adamson et al., 2014). Their relative locations – with respect to other structures and each other – have also been used as a biomarker (Vermeulen et al., 2022).

1.1 Comprehensive integrated approach

Researchers who investigate corpus callosum changes in aging, disease, and intervention studies require an accurate tool for deriving markers of corpus callosum thickness, shape, area, and volume. Instead of focusing on a single task, our framework FastSurfer-CC integrates the following five sub-tasks into a comprehensive pipeline (see Figure 4):

1. Find the mid-sagittal plane,
2. localize the anterior and posterior commissures,
3. segment CC and fornix,
4. derive morphometric estimates for CC area, length, curvature, and thickness, and
5. sub-divide the CC with a meaningful geometric sub-segmentation method.

Finding an accurate mid-sagittal plane is motivated by

the lack of clear CC boundaries in the lateral direction, often addressed by segmenting the CC in mid-sagittal slices of the brain within a pre-defined width (C. Adamson et al., 2014; B. Ardekani et al., 2013; Fischl, 2012). Determining the mid-sagittal slices accurately and independently of the image slicing direction is essential, as the shape, thickness, and area of the corpus callosum can differ considerably depending on the position and orientation of the selected mid-sagittal plane (see Figure 2). The second step, anterior- and posterior commissure localization, is required for multiple tasks: standardizing the head position in the remaining direction (nodding), finding the corpus callosum end-points, and estimating the CC location in the mid-sagittal slice as initialization for the segmentation task. Following the segmentation, the analysis can be run independently of the images’ voxel grid by converting the segmentation mask into a triangle mesh with sub-voxel resolution along the boundary. In the fourth step, building on the geometric representation of the CC, shape summary metrics like area, principal length, curvature, and local thickness can be calculated reliably. Finally, a novel and improved geometric sub-segmentation is enabled that adapts to the curved shape of the CC naturally.

1.2 Related work

Existing literature frequently reduces the challenge of CC morphometry to one of the aforementioned tasks, without integrating or evaluating the effects of other pipeline steps (Chandra et al., 2023; Gajawelli et al., 2024; Jlassi et al., 2024). We also target superior accuracy for each individual component and therefore benchmark components in isolation against specialized tools. However, jointly implementing and validating our novel and comprehensive framework (shown in Figure 4) avoids incompatibilities, e.g., a segmentation tool only working on incorrectly or poorly chosen mid-sagittal planes; accurate thickness estimation, but on low-quality segmentation maps; or chaining tools with large computational over-

head due to inconsistent pre- and postprocessing steps in the pipeline.

1.2.1 Mid-sagittal plane positioning

Finding a mid-sagittal plane for corpus callosum segmentation can be based on a template, global or local symmetry, or anatomical landmarks. Sometimes the selection of mid-sagittal slices from the image (without re-slicing) has also been formulated as a classification task. This has been achieved manually (Platten et al., 2020) and with a convolutional neural network (CNN) classifier (Brusini et al., 2022). Selecting existing slices from the image, however, oversimplifies the problem, as it disregards the left-right head rotation and tilting and might cause biases in downstream analysis. For example, tighter padding as applied in tremor cases, e.g., from MS, causes the head to be more upright compared to controls, where the head positioning is less restrained, eventually leading to bias in group comparisons as outlined in Figure 2. To correct for these effects, rotating and re-slicing of the image is necessary. Registration-based approaches, e.g., to MNI space (C. Adamson et al., 2014; Gajawelli et al., 2024; W. Huang et al., 2021) can standardize translation and rotation, thereby finding a mid-sagittal plane from the atlas space. FreeSurfer’s (Fischl, 2012) *mri_cc* tool uses the symmetry of the whole brain segmentation as an initial estimate. It employs multiple heuristics to find the mid-sagittal plane, among them an optimization step rotating the plane and maximizing an alignment score based on the left and right cerebral white matter. The *Yuki* software package (B. Ardekani et al., 2013; B. A. Ardekani & Alzheimer’s Disease Neuroimaging Initiative, 2022) also relies on registration but refines the mid-sagittal plane based on 8 landmarks, while *CCSegThickness* (short CC-Seg, C. Adamson et al., 2014) incorporates a registration with FLIRT (Jenkinson et al., 2012) to an MNI template.

Thus far, a joint evaluation and comparison of these methods is missing, and it is currently unclear which general approach is superior for corpus callosum segmenta-

tion. We aim to close this gap in our study with a rater comparison.

1.2.2 Corpus callosum and fornix segmentation

The two predominant approaches for medical image segmentation are deep learning and registration methods. For segmentation of the CC, atlas-based approaches are employed by Yuki (B. Ardekani et al., 2013; B. A. Ardekani & Alzheimer’s Disease Neuroimaging Initiative, 2022) and CCSeg (C. Adamson et al., 2014). These methods often require manual cleanup (C. Adamson et al., 2014; Raaf & Westerhausen, 2023) due to “misclassifications of pericallosal veins or the fornix” (Raaf & Westerhausen, 2023). FreeSurfer’s *mri_cc* determines an initial estimate based on the white matter of a pre-existing whole brain segmentation and then uses thresholding and other post-processing steps to segment the corpus callosum. Van Schependorn and colleagues (Van Schependorn et al., 2016) propose an active shape model combined with an atlas registration for an initial estimate. Unfortunately, the method is not easily reproducible with no released implementation or technical documentation. Multiple deep learning-based segmentation methods have been proposed. Among them are a classical U-Net (Platten et al., 2020), a modified residual attention U-Net (Chandra et al., 2023), a combination of Bi-Directional Convolutional LSTMs with a U-Net (Wong et al., 2023) for multi-modal segmentation, and a probabilistic neural network (Jlassi et al., 2024). Generally, machine learning tools have been found to be more accurate in corpus callosum segmentation (Cover et al., 2018), however, the present methods were not compared with established tools, and trained models are not available, which greatly reduces usability. To close this gap, we contribute a novel, robust, well-validated, open-source CC segmentation method.

For the segmentation of the fornix, similar registration-based (Chang et al., 2022; Fischl, 2012) and deep learning tools (Greve et al., 2021) exist. FreeSurfer (Fischl,

2012) contains two tools for fornix segmentation. *mri_cc* performs joint, corpus callosum and fornix segmentation in the mid-sagittal plane, while *ScLimbic* (Greve et al., 2021) segments 5 structures of the limbic system (hypothalamus, nucleus accumbens, fornix, basal forebrain, and septal nuclei) using a standard U-Net. *ScLimbic* segments the whole fornix and is not limited to the mid-sagittal slices – allowing for segmentation of the anterior pillars and the fimbria. The same is true for the registration-based Open-Source Hypothalamic-Fornix (OSHy-X) Atlases and Segmentation Tool (Chang et al., 2022).

1.2.3 Corpus callosum morphometry

While the area of the corpus callosum cross-section is already a valuable biomarker, atrophy in specific regions is associated with different diseases and of broad research interest. Therefore, areas of local sub-segments and thickness across the corpus callosum are highly relevant morphometrics. A plethora of sub-division schemes have been proposed, some of them based on extensive in-vivo and ex-vivo analysis. Friedrich and colleagues (Friedrich et al., 2020) give an overview on the different approaches for sub-division. The most practical methods for structural MRI are geometric approaches, which allow for a robust and generalizing sub-division. Even when considering only geometric sub-divisions applicable to structural MRI, a number of schemes have been proposed (illustrated in Figure 3):

1. The Wittelson scheme (Wittelson, 1985) defines geometric sub-divisions based on lines orthogonal to an *anchor line* through the most anterior and posterior points of the CC.
2. Jäncke et al., 1997 improve Wittelsons sub-division, by drawing the anchor line through the anterior and posterior commissures, which is more robust to segmentation errors.
3. The Hofer-Frahm scheme (Hofer & Frahm, 2006) also

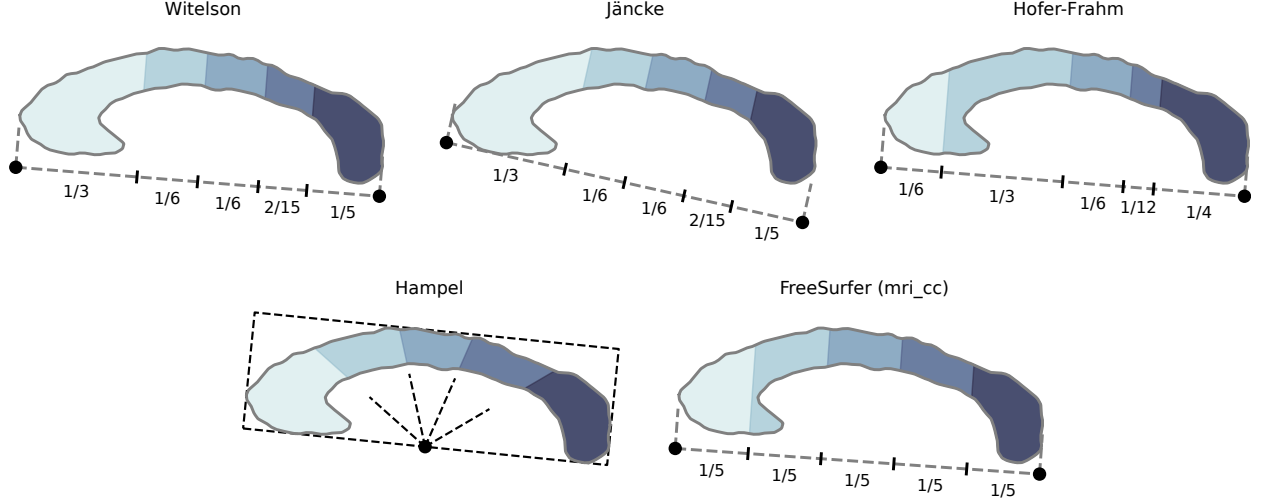


Figure 3: Overview of previously proposed sub-division schemes. To the left is the anatomical anterior.

expands on the Witelson scheme and re-defines the widths of each segment according to findings from diffusion MRI (dMRI) studies (while keeping the same anchor line).

4. Hampel et al., 1998 define an independent sub-division scheme based on equally spaced rays originating from a midpoint on the inferior border of a rectangle fitted around the CC, which yields subdivisions roughly orthogonal to the CC direction.
5. FreeSurfer (*mri_cc*) (Fischl, 2012) uses equally spaced divisions along the primary eigendirection.

Most of the proposed schemes sub-divide the CC along the inferior-superior direction. If the anterior segments are small, this can lead to unintended merging of the rostrum and body of the CC, as seen for *mri_cc* and Hofer-Frahm in Figure 3. To address this shortcoming, while preserving well-motivated sub-division fractions, we propose a shape-aware sub-division method. This method, similar to Hampel et al., 1998, cuts the sub-segments perpendicular to the CC principal direction.

Beyond area-based analysis of the corpus callosum, prior works have proposed the measurements of overall

thickness (Platten et al., 2020) and thickness profiles (C. Adamson et al., 2014; C. L. Adamson et al., 2011; Fraize et al., 2023; Lee et al., 2014; Van Schependom et al., 2016), which can be complemented with shape information (B. Ardekani et al., 2013; W. Huang et al., 2021; Joshi et al., 2013; Van Schependom et al., 2016) and sometimes also surface statistics (Gajawelli et al., 2024). Summary metrics, like mid-sagittal CC area, circularity (B. Ardekani et al., 2013; Van Schependom et al., 2016), bending angle (Platten et al., 2020), and the corpus callosum index (Figueira et al., 2007) have also been shown to be discriminative (B. Ardekani et al., 2013; Fujimori et al., 2020; Platten et al., 2020; Van Schependom et al., 2016), reliable and repeatable (Van Schependom et al., 2016) and interpretable clinical markers.

1.3 Contributions

In this work we contribute to and combine each of the previously discussed sub-tasks of i) head-pose standardization for mid-sagittal plane positioning, ii) CC & FN segmentation, iii) down-stream morphometry, and iv) sub-division. We provide a comprehensive open-source framework (FastSurfer-CC, see Figure 4) as a fast, robust, ac-

curate, and comprehensive tool for advanced corpus callosum morphometry. FastSurfer-CC:

1. Accurately identifies the optimal mid-sagittal plane tailored for corpus callosum analysis.
2. Outperforms existing specialized tools in localizing the anterior- and posterior commissures, as well as segmenting the corpus callosum and fornix.
3. Demonstrates superior robustness and accuracy on challenging cases compared to other methods.
4. Introduces novel, reliable metrics for corpus callosum length and curvature.
5. Provides efficient, precise, and robust thickness estimation along the corpus callosum.
6. Implements a new shape-aware sub-division approach compatible with previously established schemes.

2 Methods

In the following section, we introduce datasets and present the individual components: mid-sagittal plane positioning, AC & PC localization, CC & fornix segmentation, thickness, curvature and length estimation and CC sub-segmentation (see Figure 4). Finally, we describe the evaluation criteria to assess the accuracy and robustness of the aforementioned components.

2.1 Datasets

We assemble two types of datasets: 1. the training, validation, and test datasets with manual labels, 2. a downstream application dataset without manual labels.

Diverse data for training and testing is critical for method generalization and usability. Here, we start from the FastSurfer training dataset compiled from 12 datasets as described by Henschel et al., 2022 and add 7T data from the group of MR-physics at the German Center for Neurodegenerative Diseases (DZNE) and cases with resection cavities from the Uniklinikum Bonn (UK-Bonn) to further increase heterogeneity. From this large corpus

of structural MRI, we select a subset of 280 T1w scans for manual annotation and comparison of mid-sagittal planes. After the labeling process and quality control, we end up with 177 labeled volumes. These are then split into training (97), validation (31), and test-set (30), and an additional test-set consisting of cases that were challenging for raters (19), which includes cases with motion artifacts, low contrast, and brain lesions. In Appendix Figure A.2 we show examples of challenging cases, and in Appendix Table A.3 we present a breakdown of the dataset split. We evaluate performance on both randomly selected and specifically challenging cases for the following reasons:

1. Researchers often require methods to capture disease effects. Diseased groups, however, can show strong atrophy or other anomalies that make processing more challenging.
2. Datasets of specific diseases can under-represent the general population variance because of study exclusion criteria or selection bias.
3. A thorough limitation analysis helps define the constraints in which a software should be used.

Additionally, we apply our method to the PREDICT-HD dataset (predict Huntington’s disease, Paulsen et al., 2008) as a downstream evaluation. Finally, we analyze group differences between Huntington’s disease patients (N=992) and healthy controls (N=276) using the proposed morphometrics. For each participant, we select their third visit and the best quality T1w image of that visit. We also test whether the down-stream metrics produced by (only) our method show group differences. The studies were approved by the ethics board of the respective institutions.

2.2 Mid-sagittal plane positioning

We find the mid-sagittal plane by registering FastSurfer’s whole brain segmentation with the segmentation of the *fsaverage* template. More specifically, we generate two point-clouds with the centroid points of all labels present

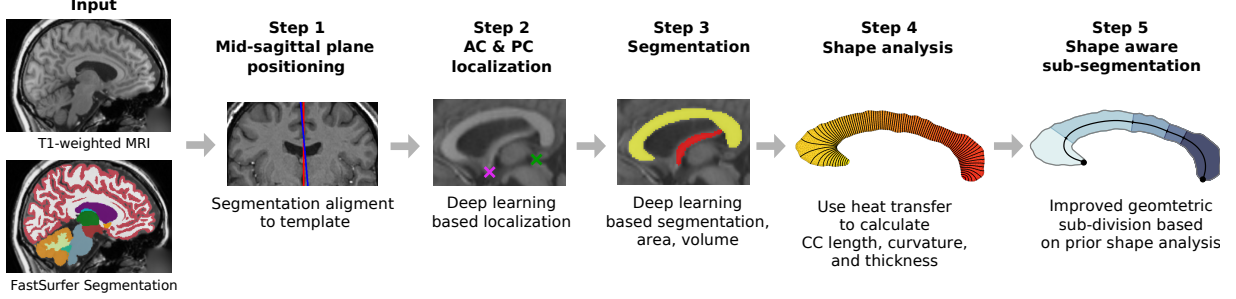


Figure 4: Overview of the proposed pipeline for corpus callosum morphometry.

in both segmentation maps. Following this, we rigidly register the point-clouds using singular value decomposition, which provides the homogeneous transformation matrix T , that maps the case segmentation space to the fsaverage template. This template is based on 40 cases from Washington University, collected by Randy Buckner and colleagues (Buckner, n.d.; FreeSurferWiki, n.d.). It is also roughly aligned to the MNI305 (Evans et al., 1993) template. The template’s mid-sagittal plane $P_{fsaverage}$ lies exactly on the middle slice in the L/R direction. Therefore, we can determine the mid-sagittal plane of the original volume P_{target} by mapping $P_{fsaverage}$ to the original images space: $P_{target} = T^{-1} \times P_{fsaverage}$.

2.3 Anterior- and posterior commissure localization

Since the mid-sagittal plane usually intersects the AC & PC points, we use the transformation T from the registration step to extract a mid-sagittal slice from the original volume for localization – effectively constraining the task to a 2D localization. To tackle this task, we train a classical DenseNet (G. Huang et al., 2016), that predicts two floating-point numbers as outputs and mean squared error as loss. These outputs determine the position of AC & PC in the image. We regularize image dimensions by re-sampling all images to 1 mm resolution and cropping the input slices to a size of 64×64 , initially centered around the third ventricle. We augment the network training by

shifting the field of view, making random contrast adjustments, adding Gaussian noise, and randomly scaling intensities. We choose the AdamW optimizer with gradient clipping. We further aim to increase the model’s accuracy and robustness using four additional strategies: i) We increase dataset variance and size by introducing a pre-training step using the FastSurfer training set (Henschel et al., 2022) with labels generated by the `acpc_detect` tool. ii) We mitigate out-of-distribution failures in case of inaccurate mid-sagittal plane positioning by training the model also on slices adjacent to the estimated optimal mid-sagittal slice. iii) We mitigate out-of-distribution errors for inaccurate fields-of-view by using an iterative inference strategy that refines the field-of-view according to the network predictions (and then runs the network again). Finally, iv) a total of 5 slices (the mid-sagittal slice plus two to either side) are used to estimate AC & PC locations and averaged.

2.4 Segmentation

Similar to AC/PC localization, we segment the CC on the mid-sagittal and adjacent slices. Here, we always consider enough slices to cover at least 2.5 mm to the left and right of the mid-sagittal plane, resulting in a standardized volumetric estimate. To aid clear separation of corpus callosum and fornix, we also segment the fornix in these mid-sagittal slices, resulting in CC and fornix outputs similar to FreeSurfer’s `mri_cc`. However, while FreeSurfer always

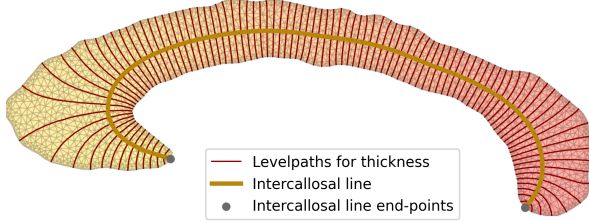


Figure 5: Final state of thickness estimation, where the intercallosal line and thickness levelpaths are calculated on the corpus callosum mesh. The solution to the Laplace solution is shown as a gradient from yellow to red.

segments and computes volumes based on only 5 slices, we increase the number of slices for sub-millimeter voxels. We then weigh the volume contribution of the first and last slice appropriately and report a corrected volume estimate that reflects a consistent width of 5 mm independent of voxel resolution. For this segmentation task, we train a variant of the FastSurferVINN (Henschel et al., 2022) architecture. This variant only operates on sagittal slices, but retains the multi-slice input (i.e., providing neighboring slices as channels) and the VINN layers (Voxel-size Independent Neural Network layers, for interpolation of the latent space). We train the network with Dice and cross-entropy loss. For data augmentation we vary the cropping of the field of view and perform random contrast changes. We use the SGD optimizer and cosine annealing with warm restarts as the learning rate scheduler.

2.5 Thickness, curvature and length estimation

Similar to CCSeg (C. L. Adamson et al., 2011), we aim to provide a method for localized morphometrics to help identify CC regions with significant thickness differences. Therefore, we develop a method to measure CC thickness with arbitrary sampling density (see Figure 5 for an example with 100 thickness measurements). We implement four changes to previous solutions: i) shape analysis with triangle meshes instead of on the voxel grid for more accu-

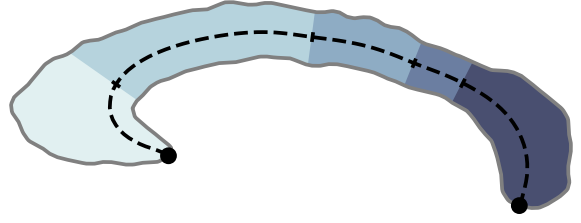


Figure 6: Proposed sub-segmentation strategy with division perpendicular to the intercallosal line. Here we use the ratios of the Hofer-Frahm sub-segmentation scheme, but other ratios are also possible. For comparison, the same corpus callosum is used as in Figure 2.

rate sub-voxel shape representation, ii) CC endpoint localization based on our reliable AC/PC landmarks even for unusual CC shapes, iii) thickness estimates via the solution to the Laplace equation (boundary conditions feature a positive and a negative charge at the opposing inferior and superior boundaries), and iv) efficient computation by estimating the level sets of the rotated Laplace gradients instead of tracing the gradient curves. Note that the zero level set to the original Laplace solution connects the CC end-points. This *intercallosal* line reliably follows the bend of the CC and can therefore be used to calculate CC curvature, length, and support different geometry-aware sub-division schemes. Level sets to the rotated solution (levelpaths for thickness in Figure 5) can be employed to similarly estimate the thickness profiles at arbitrary sampling points along the intercallosal line. We provide a detailed description of our method in Appendix A.2.

2.6 Sub-segmentation

While most previously proposed sub-segmentation schemes are carefully derived based on ex-vivo and in-vivo analysis of corpus callosum morphology and connectivity, their geometric definitions may not always produce sub-segments consistent with the intended parcellation. In Figure 3, for example, we can observe that the Hofer-Frahm sub-segmentation scheme does not separate CC

genu and body as originally intended (Hofer & Frahm, 2006), instead merging the rostrum and anterior fornix body into one sub-region. Other schemes show similar shortcomings due to the sub-optimal slicing angles with respect to the CC shape.

To address the limitations of previous CC sub-segmentation schemes, we propose a novel and robust approach for geometric sub-segmentation: we divide the CC perpendicular to the intercallosal line. In Figure 6, we apply the anatomically motivated fractions from the Hofer-Frahm scheme along the principal direction of the CC shape. This approach leads to the intended division between the posterior and superior genu. We provide a customizable, parametric implementation for all the geometric sub-segmentation schemes shown in this paper, which allows researchers to redefine sub-segmentation – e.g. according to thickness changes derived by our tool.

2.7 Implementation details

The FastSurfer-CC framework combines all of the previously discussed method steps and combines them into an efficient pipeline with the following computational improvements for speed and accuracy:

1. We interpolate only required slices into the mid-sagittal space for AC & PC localization and CC segmentation.
2. We initialize the iterative AC & PC localization using the third ventricle label.
3. We use the final AC & PC locations to crop a patch based on the AC & PC localization for the segmentation network, then we map the soft-labels (probability maps) and map them back to the original space before thresholding.
4. We calculate CC circularity and the CC index, which are previously validated CC shape descriptors (see Appendix A.3).
5. We write outputs asynchronously during processing.

The pipeline’s outputs can then be used for various tasks, including i) volumetric analysis of the corpus callosum, ii) inclusion of corpus callosum metrics in statistical analysis, and iii) standardization of head pose and position (AC at the origin, PC placed on the anterior-posterior axis and the CC centered on the left-right axis). All outputs are calculated in less than 10 seconds (Intel Xeon W-2245, 64 GB RAM, Nvidia Titan Xp, solid state drive, 256³ volume) or 40 seconds on a MacBook Pro (2.3 GHz Quad-Core Intel Core i7, 16 GB RAM, solid state drive, same volume) via docker containerization – despite the lack of a dedicated GPU.

2.8 Validation metrics

2.8.1 Manual comparison of mid-sagittal plane candidates

With the goal of identifying the best method for mid-sagittal plane positioning, we designed a custom rating tool based on Freeview – FreeSurfer’s visualization user interface (Fischl, 2012). Our tool converts a homogeneous transformation into a plane (see Section 2.2) that can be displayed independently of the volume’s voxel grid. This enables assessing whether the plane cuts the corpus callosum along its thickest lateral point while preserving local symmetry – especially the anterior and posterior commissures.

In a direct method comparison, two blinded experts compared forty mid-sagittal planes generated by each method. Raters were asked to choose the higher quality plane according to the aforementioned criteria. We determined the preferred method (i.e. method with more votes) for each rater individually. For each comparison, we selected twenty challenging plane pairs based on the lowest method agreement and twenty pairs at random. Method agreement was estimated by the average distance between planes within the brain volume. We approximated this using a cylinder centered at the RAS coordinate origin and oriented along the two planes’ average normal vector. The difference between planes was then defined as

the volume enclosed between them within the cylinder.

2.8.2 Localization comparison

We compare estimates of AC & PC location with the manually marked locations using the Euclidean distance between prediction and reference standard. Since we pay increased attention to method robustness, we show both mean and median distances. Opposed to the mean, the median is not sensitive to outliers, and the difference between the two measurements can therefore indicate whether a method produces large errors in some cases.

2.8.3 Segmentation quality

To evaluate segmentation performance with respect to the manual reference standard, we use the Dice Similarity Coefficient (DSC) and the Hausdorff Distance (HD). The DSC is defined as

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

This metric quantifies the overlap between two binary masks X and Y , which in our case represent the location and extent of brain structures. A DSC of zero indicates no overlap between the prediction and the reference standard, while a perfect match corresponds to a DSC of one. The HD is defined as

$$HD(X, Y) = \max \left\{ \sup_{x \in X} d(x, Y), \sup_{y \in Y} d(X, y) \right\} . \quad (2)$$

For the same binary masks X and Y measures the distance of the boundaries of the two structures. To provide a more robust evaluation, we use the 95th percentile of the distance rather than the maximum possible value. The HD is reported in millimeters, where 0 mm signifies a perfect match of the structures up to the 95th percentile. To determine whether one method significantly outperforms another, we apply the Wilcoxon rank-sum test, as implemented in the SciPy (Virtanen et al., 2020) library. The null hypothesis for this test assumes that the method rankings are random. To enable a fair, isolated

comparison, we use the same mid-sagittal plane estimate (by *mri.cc*) for all methods.

2.8.4 Comparison of group difference for Corpus callosum thickness

Finally, we perform a whole pipeline comparison, where we aim to find group differences in corpus callosum thickness between FastSurfer-CC and the only other method capable of generating thickness profiles: CCSeg. Here the combination of mid-sagittal plane selection, segmentation, and thickness analysis all contribute to the accurate final analysis. After running both methods, we discard all cases where either method fails (did not produce outputs) and extract the CC thickness profiles. For each of the 100 thickness values and for each method, we create a linear model that predicts thickness with age, sex, and total brain volume (from FastSurfer) as covariates. The resulting p-values are then corrected with the Benjamini and Hochberg procedure for multiple comparisons (Benjamini & Hochberg, 1995) and mapped onto the level sets of a template corpus callosum contour. The CC shape is colored according to the values of the level sets, and values between level sets are interpolated. By combining multiple CC outlines, we also generate a 3D view of the Corpus Callosum, which we use to display absolute differences between groups. Finally, we use the linear modeling with covariates from above to create separate models for each of the CC measures derived by our method and investigate which measures best explain group differences between patients and controls. We use the SciPy (Virtanen et al., 2020) and statsmodels (Seabold & Perktold, 2010) libraries for statistical testing.

3 Results

First, we evaluate methods for mid-sagittal plane positioning. Then, we determine the accuracy of our trained AC & PC localization and CC & FN segmentation models and benchmark them against state-of-the-art approaches using manually annotated cases. Finally, we evaluate the

	Rater 1		Rater 2	
SOTA Method	SOTA	FS-CC	SOTA	FS-CC
robust_register	7	12	3	13
mri_cc	6	4	10	12
Yuki	4	12	10	22
CCSeg	7	15	12	22

Table 1: Rater comparisons of the proposed method for positioning the mid-sagittal plane (FS-CC) against four state-of-the-art (SOTA) methods. The table shows votes from two independent raters in a direct comparison. The preferred method for each rater is highlighted in green.

Method	Test set random ($N = 29$)				Test set challenging ($N = 19$)			
	AC		PC		AC		PC	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
FastSurfer-CC	1.10	0.90	0.91	0.65	1.35	0.97	0.92	0.60
acpc_detect	4.77	0.71	5.54	0.65	8.84	0.71	8.51	0.76
SyN (ANTs)	2.49	1.84	2.08	1.70	3.23	2.46	2.84	1.94

Table 2: Mean and median localization error for anterior & posterior commissure on the random and challenging test sets.

full framework on the PREDICT-HD dataset, where we use all derived down-stream measures for statistical analysis.

3.1 Mid-sagittal plane positioning

We compare six methods for mid-sagittal plane positioning: i) finding the left-right symmetry axis by mid-space registration of FreeSurfer’s *mri_robust_register* (Reuter et al., 2010) to a left-right flipped version of the same volume, ii) FreeSurfer’s *mri_cc*, iii) Yuki, iv) CCSeg, iv) to FastSurfer-CC (our method). The results in Table 1 show that FastSurfer-CC outperforms all methods except for *mri_cc*. Here, method rankings disagree between raters. However, FastSurfer-CC is more than an order of magnitude faster than *mri_cc*.

3.2 Anterior- and posterior commissure localization

We compare the localization accuracy of three methods (Table 2): i) *acpc_detect* (part of the ART toolbox and the Yuki tool; B. A. Ardekani and Bachman, 2009), ii) non-linear multi-template registration using SyN (pro-

posed and validated by Pallavaram et al., 2009, Liu and Dawant, 2014, part of the ANTs toolbox; Avants et al., 2008; Tustison et al., 2021), iii) FastSurfer-CC (ours). We observe that *acpc_detect* is often highly accurate (<1 mm median error), but it produces 5 cases (10%) with extreme failures (>10 mm error, for both AC & PC). Errors of this magnitude would prevent further processing for many tasks and may introduce processing bias. Registration with SyN does not show failure cases of this magnitude but generally under-performs our method. As expected, localization methods perform much worse on the challenging test-set. Our method is a notable exception to that rule. Overall, we conclude that our method outperforms the other approaches, especially with respect to method robustness.

3.3 Segmentation

We compare four segmentation methods for CC and FN each. For the CC, we compare i) FastSurfer-CC (Ours), ii) CC.Seg, iii) Yuki, and iv) *mri_cc*. For the FN, we compare i) FastSurfer-CC (Ours), ii) SCLimbic, iii) *mri_cc*, and iv) OSHy-X. For all evaluations, we present the Dice similarity coefficient (DSC) with respect to manual seg-

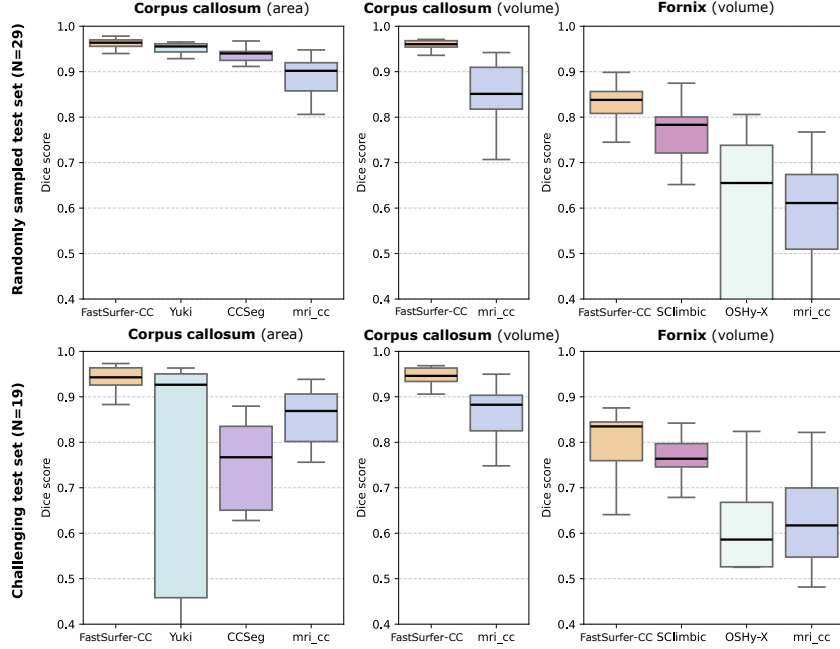


Figure 7: Dice similarity coefficient for different evaluation scenarios. All differences to our method are significant ($p < 0.005$)

mentation in Figure 7. Additionally, we compare the Hausdorff distances – where notable – in the text. The CC segmentation analysis is further sub-divided into 1. area (only the mid-sagittal plane) and 2. volume (only the 5 mid-sagittal slices segmented by *mri_cc*). Note that only FastSurfer-CC and FreeSurfer’s *mri_cc* support multi-slice evaluations.

Considering the CC segmentation, FastSurfer-CC (our method) significantly outperforms ($p < 0.01$) other methods in both the randomly selected and challenging test scenarios. FastSurfer-CC accuracy also only decreases slightly when using more challenging data – especially compared to other methods. While significantly worse than FastSurfer-CC in DSC, CCSeg narrowly outperforms FastSurfer-CC in Hausdorff distances on the random test-set (mean CCSeg 0.77, mean FastSurfer-CC 0.99 mm, median 1 mm for both, $p < 0.05$). However, on the challenging test-set CCSeg performs worst of all methods with a high Hausdorff Distance driven by out-

liers, while our method outperforms all others, retaining similar accuracy as seen on the randomly selected cases (mean CCSeg 13.34 mm, mean FastSurfer-CC 1.29 mm, median 1.41 mm and 1.0 mm respectively $p = 0.05$).

For FN segmentation, FastSurfer-CC also outperforms all other methods significantly ($p < 0.01$) for both randomly selected and challenging test scenarios. As for CC segmentation, the segmentation is remarkably reliable even in challenging scenarios with average DSC > 0.8. Finally, our method outperforms SClimbic (specialized to fornix) on both test-set and difficult test-set for DSC, but the situation is reversed for Hausdorff distances (mean SClimbic 1.63 mm, mean FastSurfer-CC 2.29 mm, median 1.41 mm and 2.03 mm respectively). Overall, we find that our method delivers superior CC & FN segmentations and is also more robust than other methods.

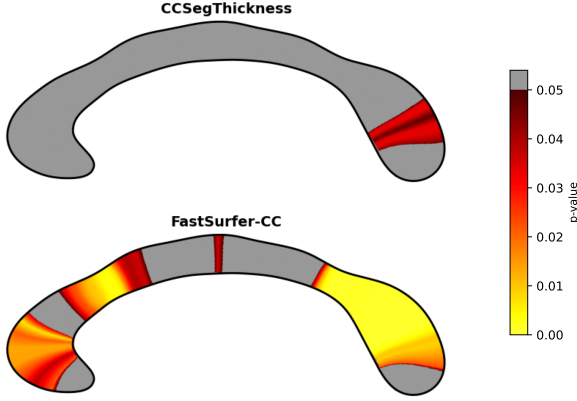


Figure 8: Corrected p-values for the group comparison of corpus callosum thickness between Huntington’s disease patients and controls mapped onto a template.

3.4 Corpus callosum morphometry in Huntington’s disease

To validate that our method can find group differences, we perform the previously introduced statistical analysis, comparing Huntington’s disease patients with controls. We use CCSeg and FastSurfer-CC methods end-to-end to generate thickness profiles (as they are the only methods with this capability) and other corpus callosum measures. We show a plot with corrected p-values in Figure 8. While only 6 of 100 thickness values show significant effects for CCSeg ($p < 0.05$, dark red in Fig. 8), our method finds significant effects in 57 of 100 values.

The Hofer-Frahm sub-segmentation yields no significant effects when using CCSeg for segmentation and post-processing ($p > 0.05$ for each sub-segment). Using FastSurfer-CC however, we find significant changes in the isthmus ($p < 0.05$, second posterior segment), which is the region that also shows the most significant changes in CC thickness (see Fig. 8). With our shape aware sub-division method (still using sub-segments defined by Hofer-Frahm) the significance of this finding increases (from $p \approx 0.04$ to $p \approx 0.01$).

Furthermore, our model provides an extensive library

of measures of corpus callosum shape. The average thickness, for example, shows the expected highly significant difference using our method ($p < 0.0001$), but not for CC-Seg ($p \approx 0.42$) using the same statistical model as above. Other measures only provided by our method also show significant differences: CC volume ($p < 0.001$), total CC perimeter ($p < 0.01$), CC circularity ($p < 0.01$), and the length of the intercallosal line ($p < 0.05$). We find no significant differences for the total CC area on the mid-sagittal plane ($p \approx 0.27$), the CC curvature (bend of the intercallosal line, $p \approx 0.39$) and the CC index ($p \approx 0.87$). We show an overview of these effects in Figure 9.

4 Discussion

Overall, we propose a fast, robust, and accurate framework for corpus callosum morphometry. We have validated its components individually and also showed that it outperforms the state-of-the-art with respect to sensitivity for a group comparison of Huntington’s disease. Furthermore, our method offers improved sub-segmentation approaches and additional downstream metrics. Besides corpus callosum analysis, our framework can be used for head pose normalization, which provides a fast alternative to rigid template registration and can aid as initialization of other affine or non-linear registration tasks. Since the alignment takes only seconds, it can be employed to pre-process large datasets and is applicable in real-time applications, e.g., for QC’ing orientation angles in diffusion MRI.

Our evaluation is specifically designed to consider the performance on unseen data of the difficult test-set to evaluate critical method robustness. This includes cases with brain lesions, strong atrophy, and artifacts, which were unseen during training. Our experiments and results demonstrate that both our method and previous state-of-the-art methods show excellent average performance on randomly sampled cases. In a preliminary setting, we established an inter-rater reliability for segmentation of 0.950 (CC DSC) and 0.852 (FN DSC). With performance

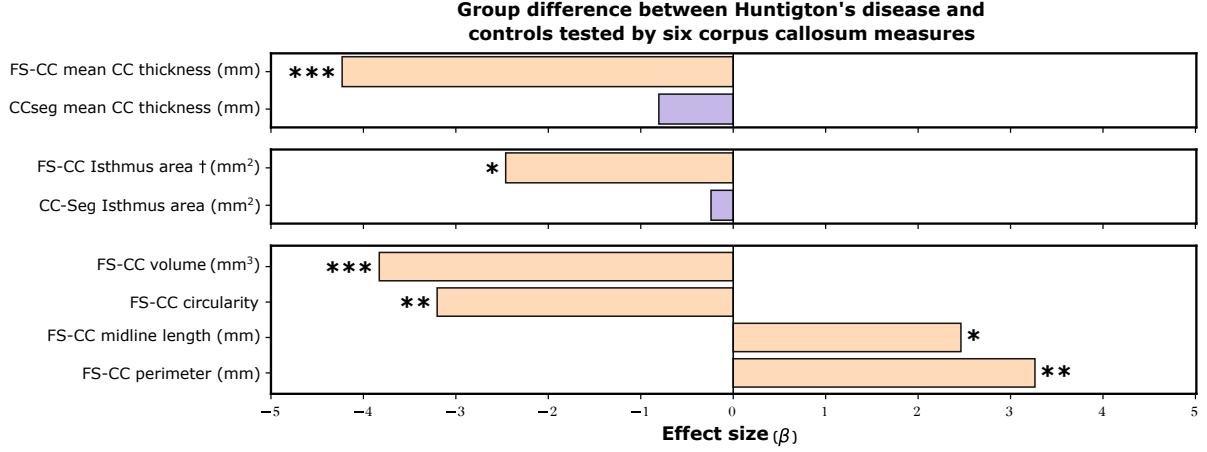


Figure 9: Effect sizes (β values in linear model), and p-values for a comparison of Huntington’s disease patients with controls. Negative values mean reduction of measure in patients. $\dagger \rightarrow$ We use FastSurfer-CC’s improved sub-division method. * $\rightarrow p < 0.05$ ** $\rightarrow p < 0.01$, *** $\rightarrow p < 0.001$, FS-CC $\rightarrow$ FastSurfer-CC

around that same level for our method, further distinction between methods is increasingly hampered by this inter-rater reliability threshold. The localization equally is highly accurate (typically less than one voxel error). The performance of the state-of-the-art methods Yuki, CCSeg, and acpc_detect is similar, yet slightly lower on high-quality data.

Nonetheless, our analysis shows that extreme errors occur regularly in competing methods – sometimes even for both challenging and randomly selected cases. While large errors can often be caught in quality control, more subtle errors may be missed and can bias down-stream analysis in populations. Therefore, we highlight that FastSurfer-CC’s accuracy remains high across the board when moving from randomly selected to challenging cases while at the same time outperforming specialized methods in good conditions. This superior accuracy and robustness leads to increased sensitivity as demonstrated in a cohort of Huntington’s disease patients and controls. In addition to the generally higher significance of our findings, we observe that the theoretical improvements to the sub-segmentation method translate to higher sensitivity in this analysis and also validate the newly introduced

summary measures like corrected CC volume, CC perimeter, and CC length measured by the intercallosal line. Nonetheless, the proposed analysis of thickness estimates has higher statistical power than area-based analysis – even when adjusting for multiple comparisons.

Our additions and contributions provide researchers with a highly sensitive, novel tool to explore and analyze CC shape changes in disease and aging. While our framework proposes new, sensitive markers of length and curvature of the CC as well as a novel geometry-aware sub-segmentation scheme, it also provides easy access to a plethora of previously proposed metrics and schemes. We will integrate FastSurfer-CC into the open-source project FastSurfer at github.com/DeepMI/FastSurfer. This update will enhance FastSurfer with rapid head pose standardization, AC & PC localization, CC & FN segmentation, CC thickness profiles, and other new morphometrics. With less than 10 seconds processing time, this contribution will also improve FastSurfer’s overall processing speed.

5 Acknowledgments

We are grateful to Dr. Jingjing Wu for her advice and support on neuroanatomy topics. We also thank Andreas Girodi for his dedication and support in the initial stages of the project. Furthermore, we thank the Rhineland Study group (Monique M.B. Breteler), the group of Toni Stöcker, and the group of Theodor Rüber for supporting the data acquisition and management.

This work is supported by DZNE institutional funds, the Federal Ministry of Research, Technology and Space of Germany (BMFTR 031L0206), by the NIH (R01-AG064027, R01 MH131586, R01 MH130899), the Chan-Zuckerberg Initiatives Essential Open Source Software for Science RFA (EOSS5 2022-252594), and by the Helmholtz Foundation Model Initiative (The Human Radiome Project). Data used in the preparation of this article were obtained in part by the OASIS Cross-Sectional with principal investigators D. Marcus, R. Buckner, J. Csernansky, J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382, and OASIS: Longitudinal: Principal Investigators: D. Marcus, R. Buckner, J. Csernansky, J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382. Further, data used in the preparation of this article were obtained from the MIRIAD database. The MIRIAD investigators did not participate in analysis or writing of this report. The MIRIAD dataset is made available through the support of the UK Alzheimer’s Society (Grant RF116). The original data collection was funded through an unrestricted educational grant from GlaxoSmithKline (Grant 6GKC). Data used in preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: <http://adni.loni.usc.edu/wp-content/uploads/>

[how_to_apply/ADNI_Acknowledgement_List.pdf](http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf). Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. Data were also provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

6 Data & code availability Statement

The source code for FastSurfer-CC will be integrated into the FastSurfer toolbox at <https://github.com/Deep-MI/FastSurfer>. The rating tool used for the user study is based on <https://github.com/deep-MI/segmentation-labeling>.

All MRI datasets, except for data of the Rhineland Study and in-house datasets, are publicly available, and references to the open-source repositories are provided in Appendix Table A.3. The UK-Bonn dataset is not publicly available since it contains information that could compromise the privacy of research participants. Access to the 7T dataset can be provided to scientists upon reasonable request to the Human MRI department of the German Center for Neurodegenerative Diseases. Data from the Rhineland Study is not publicly available because of data protection regulations; however, access can be provided to scientists in accordance with the Rhineland Study’s Data Use and Access Policy. Requests to access the data should be directed to Dr. Monique Breteler at RS-DUAC@dzne.de.

8 Declaration of competing interests

The authors do not declare any competing interests.

7 Author contributions

Clemens Pollak: Methodology, Conceptualization, Formal analysis, Investigation, Data Curation, Software, Validation, Writing – original draft, Writing – review & editing, Visualization.

Kersten Diers: Conceptualization, Data Curation, Resources, Writing – review & editing.

Santiago Estrada: Data Curation, Resources, Writing – review & editing, Visualization.

David Kügler: Conceptualization, Writing – original draft, Writing – review & editing, Supervision

Martin Reuter: Conceptualization, Methodology, Software, Resources, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition.

A Appendix

A.1 Data description

For method development, we manually labeled 177 cases with corpus callosum, fornix, anterior and posterior commissures, and split them into training, validation, and test-set. We also established another difficult test-set with cases that are especially challenging due to brain lesions, motion artifacts, strong atrophy, and low imaging contrast. For labeling, we use mid-sagittal planes generated by FreeSurfer’s *mri_cc* tool. Since this FreeSurfer tool uses whole brain segmentations, we quality control the whole brain segmentations prior to processing and use lesion inpainting (Pollak et al., 2025) when large areas of damaged tissues are present to circumvent errors. We show an overview of the datasets used for annotation in Table A.3.

A.2 Calculating corpus callosum thickness, length and curvature

To derive accurate morphometrics from the CC segmentation, we first detach the corpus callosum contour from the voxel-grid by applying a Gaussian filter to the binary label. Next, we apply the marching squares algorithm implemented in the scikit-image library (van der Walt et al., 2014). Next, the closed boundary contour is triangulated with the meshpy library (Steinbrecher & Popp, 2021), providing a high-quality, dense triangle mesh of the CC shape (see Fig. 5). To split the CC boundary into an inferior and superior curve at its end-points, we use the relative position of the anterior and posterior commissures. In contrast to previous heuristics (e.g., using the curvature of the contour, C. L. Adamson et al., 2011), the AC & PC are better references because they are deformed together with the CC, leading to better intercallosal lines in brains with strong atrophy or other disease-related changes. Based on the training set, we design a heuristic that gives visually accurate endpoints by first finding intermediate anchor points with respect to the

AC-PC line and then selecting the closest points on the contour as anterior and posterior endpoints. After deriving the end-points, we split the contour into an upper and lower portion, apply appropriate boundary conditions (of $f_{\text{inf/sup}} = -1/1$ along the inferior and superior boundary and $f_{\text{ep}} = 0$ at the two endpoints), and solve the Laplace equation $\Delta f = 0$ for all interior points on the mesh. We extend the LaPy Python library¹ (Reuter et al., 2006; Wachinger et al., 2015) for these and the following operations. We extract the zero-level set of the Laplace solution as a piecewise linear path to obtain an *intercallosal line*, running from one endpoint to the other. We then resample the intercallosal line to have exactly $N + 2$ equidistant points (N being the number of the desired interior thickness estimates, default $N = 100$). Finally, to avoid tracing gradients for thickness estimates, we rotate the Laplace solution on the mesh so that the new level sets are orthogonal to the level sets of the original solution. This is done by rotating the gradients around the normals by 90 degrees, then solving the Poisson equation $\Delta f = h$ with the divergence of the rotated gradient as h . It is now easy to extract the level set curves at the N locations along the intercallosal line. The individual lengths of these level set curves provide the thickness profile of the CC. Furthermore, total CC length and curvature are computed directly from the intercallosal line and are also provided.

A.3 Additional corpus callosum metrics

In addition to thickness profiles, curvature, midline length, and sub-segmentation areas, we also include two previously proposed metrics for corpus callosum analysis and describe them here for documentation purposes.

A.3.1 Circularity

Circularity is a low-descriptive shape measure comparing area A to boundary length L , which is maximized by the

¹<https://github.com/deep-mi/LaPy>

Dataset	# Train	# Val	# Test	# Difficult	# Total
ABIDE (Di Martino et al., 2014)	8	2	4	0	14
ABIDE-II (Di Martino et al., 2017)	3	1	0	0	4
ADNI (Jack Jr et al., 2008)	15	7	3	6	32
LA5C (Poldrack et al., 2016)	10	5	0	2	17
MIRIAD (Malone et al., 2013)	11	2	3	1	17
OAS1 (Marcus et al., 2007)	8	4	5	1	18
OAS2 (Marcus et al., 2010)	8	2	2	2	14
Rhineland Study (1 mm (Koch et al., 2024)	13	2	4	0	20
Rhineland Study (Koch et al., 2024)	9	3	7	1	20
HCP (Bookheimer et al., 2019; Van Essen et al., 2012)	5	3	0	1	9
IXI (“IXI – Information eXtraction from Images”, n.d.)	0	0	0	2	2
7T (in-house)	7	0	2	0	9
UK-Bonn epilepsy (in-house)	0	0	0	3	3
Combined	97	31	30	19	177

Table A.3: Overview of the used datasets and their inclusion in the sub-sets: training, validation, test, difficult test.

disc, where it becomes 1:

$$Circ = 4\pi \frac{A}{L^2}. \quad (3)$$

The corpus callosum circularity was proposed by (Van Schependorn et al., 2018) as a marker for Alzheimer’s disease. CC atrophy and thinning from aging or disease will decrease area faster than boundary and therefore lead to a decreasing circularity. While our framework offers more fine-grained measures of length, thickness, and curvature individually, some researchers may desire an aggregated metric, which is why we include the corpus callosum circularity in the pipeline output.

A.3.2 Corpus callosum index

Similar to the corpus callosum circularity, the corpus callosum index is also an aggregate metric that tries to condense the complex shape changes of the CC by aggregating a few straightforward measurements that traditionally were easy to obtain manually. It was first proposed as a proxy for brain volume in the analysis of multiple sclerosis (MS) patients (Yaldizli et al., 2010). We have outlined the calculation in Figure A.1. Below are the detailed steps to calculate the index from a corpus callosum contour:

1. Draw the longest possible line between the anterior

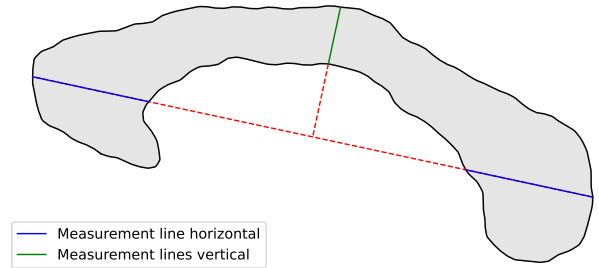


Figure A.1: The corpus callosum index is defined as the sum of the two horizontal measurement segments (blue) and the vertical measurement segment (green).

- posterior edges of the CC (cutting the CC into 3 parts)
2. Draw a second line perpendicular to the first and crossing the first lines midpoint (cutting the CC in the middle, into overall 4 parts)
3. Measure the thickness of cuts at the three cutting locations and sum them.

A.4 Difficult examples

In Figure A.2, we show examples from the difficult test set along with the rater annotation and network output. In case A the contrast of corpus callosum and surroundings is low (note that we use standardized automated contrast

adjustment by freeview for all figures). In case B strong artifacts are present (see image background for a clear view of ringing artifacts). In cases B, C, and D, the corpus callosum is thin and deformed, making segmentation challenging.

References

- Adamson, C., Beare, R., Walterfang, M., & Seal, M. (2014). Software pipeline for midsagittal corpus callosum thickness profile processing: Automated segmentation, manual editor, thickness profile generator, group-wise statistical comparison and results display. *Neuroinformatics*, 12, 595–614.
- Adamson, C. L., Wood, A. G., Chen, J., Barton, S., Reutens, D. C., Pantelis, C., Velakoulis, D., & Walterfang, M. (2011). Thickness profile generation for the corpus callosum using laplace’s equation. *Human Brain Mapping*, 32(12), 2131–2140.
- Ardekani, B., Lee, S., & Bachman, A. (2013). Automated segmentation of corpus callosum in mri scans: Application to tracking alzheimer’s disease progression. *Conference Abstract: ACNS-2013 Australasian Cognitive Neuroscience Society Conference*.
- Ardekani, B. A., & Alzheimer’s Disease Neuroimaging Initiative. (2022). A new approach to symmetric registration of longitudinal structural mri of the human brain. *Journal of Neuroscience Methods*, 373, 109563.
- Ardekani, B. A., & Bachman, A. H. (2009). Model-based automatic detection of the anterior and posterior commissures on mri scans. *Neuroimage*, 46(3), 677–682.
- Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1), 26–41.
- Baudo, M. C., Colombo, E. E., Pérez, M. A., Rahn, S. I., Domenech, N. C., Chuang, J., Forlizzi, V., Buzosi, P., Baldoncini, M., & Bendersky, M. (2022). Three-dimensional anatomy of the anterior commissure: A tractography and anatomical study. *World neurosurgery*, 159, e365–e374.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1), 289–300.
- Bookheimer, S. Y., Salat, D. H., Terpstra, M., Ances, B. M., Barch, D. M., Buckner, R. L., Burgess, G. C., Curtiss, S. W., Diaz-Santos, M., Elam, J. S., et al. (2019). The lifespan human connectome project in aging: An overview. *Neuroimage*, 185, 335–348.
- Brusini, I., Platten, M., Ouellette, R., Piehl, F., Wang, C., & Granberg, T. (2022). Automatic deep learning multicontrast corpus callosum segmentation in multiple sclerosis. *Journal of Neuroimaging*, 32(3), 459–470.
- Buckner, R. (n.d.). <https://surfer.nmr.mgh.harvard.edu/fswiki/Buckner40Adni60Testing>
- Çavdar, S., Aydın, A. E., Algin, O., & Aydın, S. (2021). The complex structure of the anterior white commissure of the human brain: Fiber dissection and tractography study. *World Neurosurgery*, 147, e111–e117.
- Chandra, A., Verma, S., Raghuvanshi, A., & Bodhey, N. K. (2023). Pccs-rau-net: Automated parcelated corpus callosum segmentation from brain mri images using modified residual attention u-net. *Biocybernetics and Biomedical Engineering*, 43(2), 403–427.
- Chang, J., Steyn, F., Ngo, S., Henderson, R., Guo, C., Bollmann, S., Fripp, J., Barth, M., & Shaw, T. (2022). Open-source hypothalamic-fornix (oshyx) atlases and segmentation tool for 3t and 7t. *Journal of Open Source Software*, 7(76), 4368.

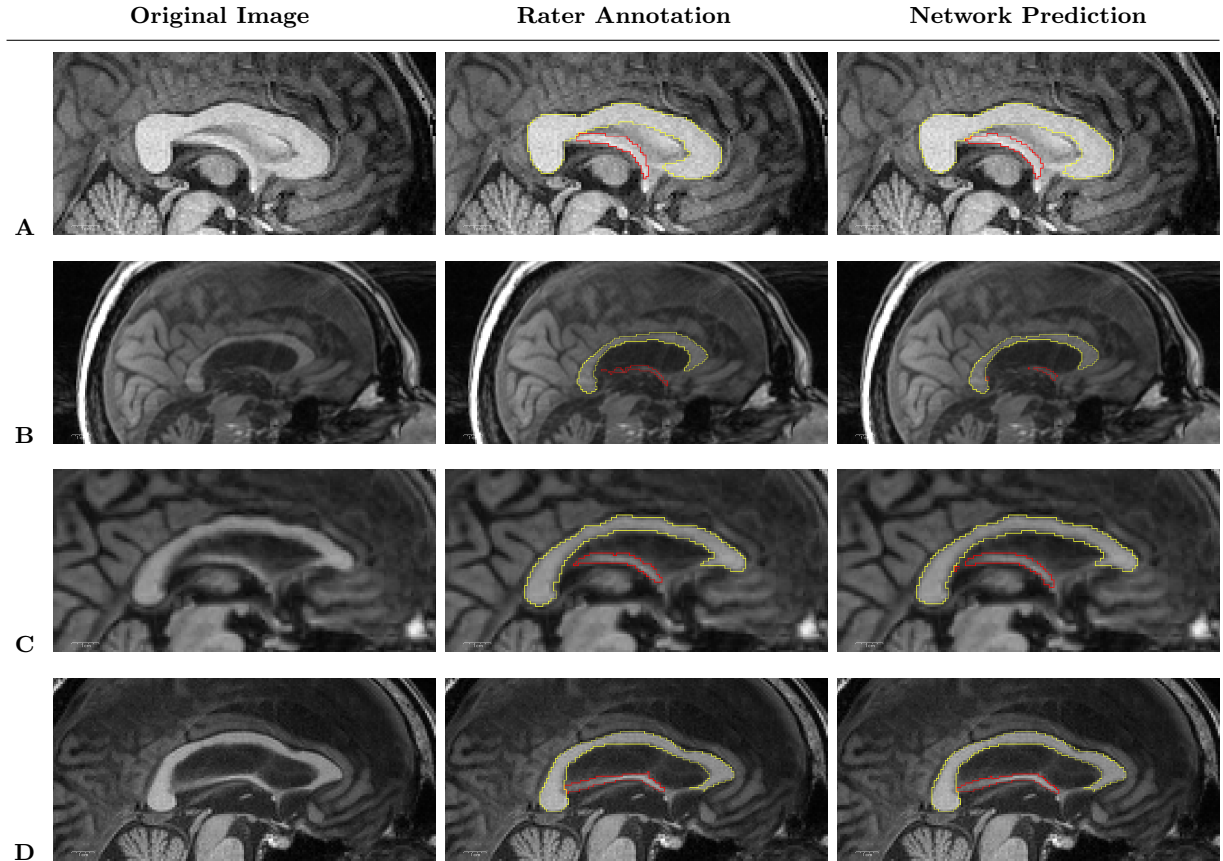


Figure A.2: Example cases from the difficult test set with rater annotation and network prediction (unseen during training).

- Cover, G., Herrera, W. G., Bento, M. P., Appenzeller, S., & Rittner, L. (2018). Computational methods for corpus callosum segmentation on mri: A systematic literature review. *Computer methods and programs in biomedicine*, 154, 25–35.
- Di Martino, A., O’connor, D., Chen, B., Alaerts, K., Anderson, J. S., Assaf, M., Balsters, J. H., Baxter, L., Beggiato, A., Bernaerts, S., et al. (2017). Enhancing studies of the connectome in autism using the autism brain imaging data exchange ii. *Scientific data*, 4(1), 1–15.
- Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M., et al. (2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6), 659–667.
- Di Paola, M., Spalletta, G., & Caltagirone, C. (2010). In vivo structural neuroanatomy of corpus callosum in alzheimer’s disease and mild cognitive impairment using different mri techniques: A review. *Journal of Alzheimer’s disease*, 20(1), 67–95.
- Evans, A. C., Collins, D. L., Mills, S., Brown, E. D., Kelly, R. L., & Peters, T. M. (1993). 3d sta-

- tistical neuroanatomical models from 305 mri volumes. *1993 IEEE conference record nuclear science symposium and medical imaging conference*, 1813–1817.
- Figueira, F. F. A., Santos, V. S. d., Figueira, G. M. A., & Silva, Â. C. M. d. (2007). Corpus callosum index: A practical method for long-term follow-up in multiple sclerosis. *Arquivos de neuro-psiquiatria*, *65*, 931–935.
- Fischl, B. (2012). Freesurfer. *Neuroimage*, *62*(2), 774–781.
- Fraize, J., Convert, G., Leprince, Y., Sylvestre-Marconville, F., Kerdreux, E., Auzias, G., Lefèvre, J., Delorme, R., Elmaleh-Bergès, M., Hertz-Pannier, L., et al. (2023). Mapping corpus callosum surface reduction in fetal alcohol spectrum disorders with sulci and connectivity-based parcellation. *Frontiers in Neuroscience*, *17*, 1188367.
- FreeSurferWiki. (n.d.). <https://surfer.nmr.mgh.harvard.edu/fswiki/FsAverage>
- Friedrich, P., Forkel, S. J., & de Schotten, M. T. (2020). Mapping the principal gradient onto the corpus callosum. *Neuroimage*, *223*, 117317.
- Fujimori, J., Uryu, K., Fujihara, K., Wattjes, M. P., Suzuki, C., & Nakashima, I. (2020). Measurements of the corpus callosum index and fractional anisotropy of the corpus callosum and their cutoff values are useful to assess global brain volume loss in multiple sclerosis. *Multiple sclerosis and related disorders*, *45*, 102388.
- Gajawelli, N., Paulli, A., Deoni, S., Paquette, N., Darakjian, D., Salazar, C., Dean, D., O’Muircheartaigh, J., Nelson, M. D., Wang, Y., et al. (2024). Surface-based morphometry of the corpus callosum in young children of ages 1–5. *Human Brain Mapping*, *45*(9), e26693.
- Grande, I., Berk, M., Birmaher, B., & Vieta, E. (2016). Bipolar disorder. *The Lancet*, *387*(10027), 1561–1572.
- Greve, D. N., Billot, B., Cordero, D., Hoopes, A., Hoffmann, M., Dalca, A. V., Fischl, B., Iglesias, J. E., & Augustinack, J. C. (2021). A deep learning toolbox for automatic segmentation of subcortical limbic structures from mri images. *Neuroimage*, *244*, 118610.
- Hampel, H., Teipel, S. J., Alexander, G. E., Horwitz, B., Teichberg, D., Schapiro, M. B., & Rapoport, S. I. (1998). Corpus callosum atrophy is a possible indicator of region- and cell type-specific neuronal degeneration in alzheimer disease: A magnetic resonance imaging analysis. *Archives of neurology*, *55*(2), 193–198.
- Henschel, L., Kügler, D., & Reuter, M. (2022). Fastsurfer-vinn: Building resolution-independence into deep learning segmentation methods—a solution for highres brain mri. *NeuroImage*, *251*, 118933.
- Hofer, S., & Frahm, J. (2006). Topography of the human corpus callosum revisited—comprehensive fiber tractography using diffusion tensor magnetic resonance imaging. *Neuroimage*, *32*(3), 989–994.
- Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2016). Densely connected convolutional networks, aug. *arXiv preprint arXiv:1608.06993*.
- Huang, W., Chen, M., Lyu, G., & Tang, X. (2021). A deformation-based shape study of the corpus callosum in first episode schizophrenia. *Frontiers in Psychiatry*, *12*, 621515.
- Ixi – information extraction from images. (n.d.). <https://brain-development.org/ixi-dataset/>
- Jaatela, J., Nurmi, T., Vallinoja, J., Mäenpää, H., Sairanen, V., & Piitulainen, H. (2023). Altered corpus callosum structure in adolescents with cerebral palsy: Connection to gait and balance. *Brain Structure and Function*, *228*(8), 1901–1915.
- Jack Jr, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L. Whitwell, J., Ward, C., et al. (2008). The alzheimer’s disease neuroimaging initiative (adni): Mri methods. *Journal of Mag-*

- netic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4), 685–691.
- Jäncke, L., Staiger, J. F., Schlaug, G., Huang, Y., & Steinmetz, H. (1997). The relationship between corpus callosum size and forebrain volume. *Cerebral cortex (New York, NY: 1991)*, 7(1), 48–56.
- Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., & Smith, S. M. (2012). Fsl. *Neuroimage*, 62(2), 782–790.
- Jlassi, A., Elbedoui, K., Barhoumi, W., & Maktouf, C. (2024). 3dcc-mpnn: Automated 3d reconstruction of corpus callosum based on modified pnn and marching cubes. *Evolving Systems*, 1–27.
- Joshi, S. H., Narr, K. L., Philips, O. R., Nuechterlein, K. H., Asarnow, R. F., Toga, A. W., & Woods, R. P. (2013). Statistical shape analysis of the corpus callosum in schizophrenia. *Neuroimage*, 64, 547–559.
- Koch, A., Stirnberg, R., Estrada, S., Zeng, W., Lohner, V., Shahid, M., Ehses, P., Pracht, E. D., Reuter, M., Stoecker, T., et al. (2024). Versatile mri acquisition and processing protocol for population-based neuroimaging. *Nature Protocols*, 1–23.
- Lacalle-Aurioles, M., & Iturria-Medina, Y. (2023). Fornix degeneration in risk factors of alzheimer’s disease, possible trigger of cognitive decline. *Cerebral Circulation-Cognition and Behavior*, 4, 100158.
- Lee, S. H., Yu, D., Bachman, A. H., Lim, J., & Ardekani, B. A. (2014). Application of fused lasso logistic regression to the study of corpus callosum thickness in early alzheimer’s disease. *Journal of neuroscience methods*, 221, 78–84.
- Liu, Y., & Dawant, B. M. (2014). Automatic detection of the anterior and posterior commissures on mri scans using regression forests. *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1505–1508.
- Luders, E., Thompson, P. M., & Toga, A. W. (2010). The development of the corpus callosum in the healthy human brain. *Journal of Neuroscience*, 30(33), 10985–10990.
- Malone, I. B., Cash, D., Ridgway, G. R., MacManus, D. G., Ourselin, S., Fox, N. C., & Schott, J. M. (2013). Miriad—public release of a multiple time point alzheimer’s mr imaging dataset. *NeuroImage*, 70, 33–36.
- Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2010). Open access series of imaging studies: Longitudinal mri data in nondemented and demented older adults. *Journal of cognitive neuroscience*, 22(12), 2677–2684.
- Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2007). Open access series of imaging studies (oasis): Cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9), 1498–1507.
- Steinbrecher, I., & Popp, A. (2021). MeshPy – A general purpose 3D beam finite element input generator. <https://imcs-compsim.github.io/meshpy>
- Nowrangi, M. A., & Rosenberg, P. B. (2015). The fornix in mild cognitive impairment and alzheimer’s disease. *Frontiers in aging neuroscience*, 7, 1.
- Ozturk, A., Smith, S., Gordon-Lipkin, E., Harrison, D., Shiee, N., Pham, D., Caffo, B., Calabresi, P., & Reich, D. (2010). Mri of the corpus callosum in multiple sclerosis: Association with disability. *Multiple Sclerosis Journal*, 16(2), 166–177.
- Pallavaram, S., Dawant, B. M., Koyama, T., Yu, H., Neimat, J., Konrad, P. E., & D’Haese, P.-F. (2009). Validation of a fully automatic method for the routine selection of the anterior and posterior commissures in magnetic resonance images. *Stereotactic and functional neurosurgery*, 87(3), 148–154.
- Paulsen, J. S., Langbehn, D. R., Stout, J. C., Aylward, E., Ross, C. A., Nance, M., Guttman, M.,

- Johnson, S., MacDonald, M., Beglinger, L. J., et al. (2008). Detection of huntington’s disease decades before diagnosis: The predict-hd study. *Journal of Neurology, Neurosurgery & Psychiatry*, 79(8), 874–880.
- Pinto, Y., de Haan, E. H., & Lamme, V. A. (2017). The split-brain phenomenon revisited: A single conscious agent with split perception. *Trends in cognitive sciences*, 21(11), 835–851.
- Piven, J., Bailey, J., Ranson, B. J., Arndt, S., et al. (1997). An mri study of the corpus callosum in autism. *American Journal of Psychiatry*, 154(8), 1051–1056.
- Platten, M., Martola, J., Fink, K., Ouellette, R., Piehl, F., & Granberg, T. (2020). Mri-based manual versus automated corpus callosum volumetric measurements in multiple sclerosis. *Journal of Neuroimaging*, 30(2), 198–204.
- Poldrack, R. A., Congdon, E., Triplett, W., Gorgolewski, K., Karlsgodt, K., Mumford, J., Sabb, F., Freimer, N., London, E., Cannon, T., et al. (2016). A phenome-wide examination of neural and cognitive function. *Scientific data*, 3(1), 1–12.
- Pollak, C., Kügler, D., Bauer, T., Rüber, T., & Reuter, M. (2025). Fastsurfer-lit: Lesion inpainting tool for whole-brain mri segmentation with tumors, cavities, and abnormalities. *Imaging Neuroscience*, 3, imag.a.00446.
- Raaf, N., & Westerhausen, R. (2023). Hand preference and the corpus callosum: Is there really no association? *Neuroimage: Reports*, 3(1), 100160.
- Raybaud, C. (2010). The corpus callosum, the other great forebrain commissures, and the septum pellucidum: Anatomy, development, and malformation. *Neuroradiology*, 52(6), 447–477.
- Reuter, M., Rosas, H. D., & Fischl, B. (2010). Highly accurate inverse consistent registration: A robust approach. *Neuroimage*, 53(4), 1181–1196.
- Reuter, M., Wolter, F.-E., & Peinecke, N. (2006). Laplace–beltrami spectra as ‘shape-dna’ of surfaces and solids. *Computer-Aided Design*, 38(4), 342–366.
- Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. *SciPy*, 7(1), 92–96.
- Tustison, N. J., Cook, P. A., Holbrook, A. J., Johnson, H. J., Muschelli, J., Devenyi, G. A., Duda, J. T., Das, S. R., Cullen, N. C., Gillen, D. L., et al. (2021). The antsx ecosystem for quantitative biological and medical imaging. *Scientific reports*, 11(1), 9068.
- Unterberger, I., Bauer, R., Walser, G., & Bauer, G. (2016). Corpus callosum and epilepsies. *Seizure*, 37, 55–60.
- van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., Yu, T., & the scikit-image contributors. (2014). Scikit-image: Image processing in Python. *PeerJ*, 2, e453. <https://doi.org/10.7717/peerj.453>
- Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S. W., et al. (2012). The human connectome project: A data acquisition perspective. *Neuroimage*, 62(4), 2222–2231.
- Van Schependom, J., Jain, S., Cambron, M., Vanbinst, A.-M., De Mey, J., Smeets, D., & Nagels, G. (2016). Reliability of measuring regional callosal atrophy in neurodegenerative diseases. *NeuroImage: Clinical*, 12, 825–831.
- Van Schependom, J., Niemantsverdriet, E., Smeets, D., & Engelborghs, S. (2018). Callosal circularity as an early marker for alzheimer’s disease. *NeuroImage: Clinical*, 19, 516–526.
- Vermeulen, C. L., du Toit, P. J., Venter, G., & Human-Baron, R. (2022). A morphological study of the shape of the corpus callosum in normal,

- schizophrenic and bipolar patients. *Journal of anatomy*, 242(2), 153–163.
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., . . . SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17, 261–272. <https://doi.org/10.1038/s41592-019-0686-2>
- Wachinger, C., Golland, P., Kremen, W., Fischl, B., Reuter, M., Initiative, A. D. N., et al. (2015). Brainprint: A discriminative characterization of brain morphology. *NeuroImage*, 109, 232–248.
- Witelson, S. F. (1985). The brain connection: The corpus callosum is larger in left-handers. *Science*, 229(4714), 665–668.
- Wong, K. K., Xu, W., Ayoub, M., Fu, Y.-L., Xu, H., Shi, R., Zhang, M., Su, F., Huang, Z., & Chen, W. (2023). Brain image segmentation of the corpus callosum by combining bi-directional convolutional lstm and u-net using multi-slice ct and mri. *Computer Methods and Programs in Biomedicine*, 238, 107602.
- Woodruff, P., McManus, I., & David, A. (1995). Meta-analysis of corpus callosum size in schizophrenia. *Journal of Neurology, Neurosurgery & Psychiatry*, 58(4), 457–461.
- Yaldizli, Ö., Atefy, R., Gass, A., Sturm, D., Glassl, S., Tetttenborn, B., & Putzki, N. (2010). Corpus callosum index and long-term disability in multiple sclerosis patients. *Journal of neurology*, 257, 1256–1264.
- Yang, K., Wu, Z., Long, J., Li, W., Wang, X., Hu, N., Zhao, X., & Sun, T. (2023). White matter changes in parkinson’s disease. *npj Parkinson’s Disease*, 9(1), 150.
- Zhuang, L., Wen, W., Trollor, J. N., Kochan, N. A., Reppermund, S., Brodaty, H., & Sachdev, P. (2012). Abnormalities of the fornix in mild cognitive impairment are related to episodic memory loss. *Journal of Alzheimer’s Disease*, 29(3), 629–639.

Bibliography

1. Merriam-Webster online dictionary. *Robustness* <https://www.merriam-webster.com/dictionary/robustness>. [Accessed 21-11-2025].
2. Pollak, C., Kügler, D., Breteler, M. M. & Reuter, M. Quantifying MR head motion in the Rhineland Study–A robust method for population cohorts. *NeuroImage* **275**, 120176 (2023).
3. Pollak, C., Kügler, D. & Reuter, M. *Estimating Head Motion from Mr-Images* in 2023 *IEEE 20th International Symposium on Biomedical Imaging (ISBI)* (2023), 1–5.
4. Pollak, C., Kügler, D., Bauer, T., Rüber, T. & Reuter, M. FastSurfer-LIT: Lesion inpainting tool for whole-brain MRI segmentation with tumors, cavities, and abnormalities. *Imaging neuroscience* **3**, imag_a_00446 (2025).
5. Pollak, C., Diers, K., Estrada, S., Kügler, D. & Reuter, M. *FastSurfer-CC: A robust, accurate, and comprehensive framework for corpus callosum morphometry* 2025. arXiv: 2511.16471 [cs.CV].
6. De Bie, H. M., Boersma, M., Wattjes, M. P., Adriaanse, S., Vermeulen, R. J., Oostrom, K. J., Huisman, J., Veltman, D. J. & Delemarre-Van de Waal, H. A. Preparing children with a mock scanner training protocol results in high quality structural and functional MRI scans. *European journal of pediatrics* **169**, 1079–1085 (2010).
7. Siddiqui, Z., Singh, P., Kushwaha, S. & Srivastava, R. MRI and fear of confined space: A cause and effect relationship. *International Journal of Contemporary Medicine Surgery and Radiology* **2**, 19–24 (2017).
8. Huijbers, W., Van Dijk, K. R., Boenniger, M. M., Stirnberg, R. & Breteler, M. M. Less head motion during MRI under task than resting-state conditions. *Neuroimage* **147**, 111–120 (2017).
9. Vanderwal, T., Kelly, C., Eilbott, J., Mayes, L. C. & Castellanos, F. X. Inscapes: A movie paradigm to improve compliance in functional magnetic resonance imaging. *Neuroimage* **122**, 222–232 (2015).
10. Greene, D. J., Koller, J. M., Hampton, J. M., Wesevich, V., Van, A. N., Nguyen, A. L., Hoyt, C. R., McIntyre, L., Earl, E. A., Klein, R. L., *et al.* Behavioral interventions for reducing head motion during MRI scans in children. *Neuroimage* **171**, 234–245 (2018).

11. Slipsager, J. M., Ellegaard, A. H., Glimberg, S. L., Paulsen, R. R., Tisdall, M. D., Wighton, P., van der Kouwe, A., Marner, L., Henriksen, O. M., Law, I., *et al.* Markerless motion tracking and correction for PET, MRI, and simultaneous PET/MRI. *Plos one* **14**, e0215524 (2019).
12. Gagliardo, C., Feraco, P., Contrino, E., D'Angelo, C., Geraci, L., Salvaggio, G., Gagliardo, A., La Grutta, L., Midiri, M. & Marrale, M. Ultra-low-field MRI: a David versus Goliath challenge in modern imaging. *La radiologia medica*, 1–18 (2025).
13. Boulant, N., Mauconduit, F., Gras, V., Amadon, A., Le Ster, C., Luong, M., Massire, A., Pallier, C., Sabatier, L., Bottlaender, M., *et al.* In vivo imaging of the human brain with the Iseult 11.7-T MRI scanner. *Nature Methods* **21**, 2013–2016 (2024).
14. Gunter, J. L., Bernstein, M. A., Borowski, B. J., Ward, C. P., Britson, P. J., Felmlee, J. P., Schuff, N., Weiner, M. & Jack, C. R. Measurement of MRI scanner performance with the ADNI phantom. *Medical physics* **36**, 2193–2205 (2009).
15. Peltonen, J. I., Mäkelä, T., Lehmonen, L., Sofiev, A. & Salli, E. Inter-and intra-scanner variations in four magnetic resonance imaging image quality parameters. *Journal of Medical Imaging* **7**, 065501–065501 (2020).
16. Brenner, D., Stirnberg, R., Pracht, E. D. & Stöcker, T. Two-dimensional accelerated MP-RAGE imaging with flexible linear reordering. *Magnetic Resonance Materials in Physics, Biology and Medicine* **27**, 455–462 (2014).
17. Çukur, T., Dar, S. U., Nezhad, V. A., Jun, Y., Kim, T. H., Fujita, S. & Bilgic, B. A tutorial on mri reconstruction: from modern methods to clinical implications. *IEEE Transactions on Biomedical Engineering* (2025).
18. Zaitsev, M., Maclaren, J. & Herbst, M. Motion artifacts in MRI: A complex problem with many partial solutions. *Journal of magnetic resonance imaging : JMRI* **42**, 887–901 (2015).
19. Reuter, M., Tisdall, M. D., Qureshi, A., Buckner, R. L., van der Kouwe, A. J. & Fischl, B. Head motion during MRI acquisition reduces gray matter volume and thickness estimates. *Neuroimage* **107**, 107–115 (2015).
20. Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., *et al.* Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* **33**, 341–355 (2002).
21. Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B. & Reuter, M. Fastsurfer-a fast and accurate deep learning based neuroimaging pipeline. *NeuroImage* **219**, 117012 (2020).

22. Tustison, N. J., Cook, P. A., Holbrook, A. J., Johnson, H. J., Muschelli, J., Devenyi, G. A., Duda, J. T., Das, S. R., Cullen, N. C., Gillen, D. L., *et al.* The ANTsX ecosystem for quantitative biological and medical imaging. *Scientific reports* **11**, 9068 (2021).
23. Fischl, B. FreeSurfer. *Neuroimage* **62**, 774–781 (2012).
24. Henschel, L., Kügler, D. & Reuter, M. FastSurferVINN: Building resolution-independence into deep learning segmentation methods—A solution for HighRes brain MRI. *NeuroImage* **251**, 118933 (2022).
25. Engelhardt, L. E., Roe, M. A., Juranek, J., DeMaster, D., Harden, K. P., Tucker-Drob, E. M. & Church, J. A. Children’s head motion during fMRI tasks is heritable and stable over time. *Developmental cognitive neuroscience* **25**, 58–68 (2017).
26. Schauman, S., van Niekerk, A., Norbeck, O., Rydén, H., Avventi, E. & Skare, S. An Exploration of Motion-Sampling Interactions in 3D MRI for Neuroimaging. *Magnetic Resonance in Medicine* (2025).
27. Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A. & Gee, J. C. N4ITK: improved N3 bias correction. *IEEE transactions on medical imaging* **29**, 1310–1320 (2010).
28. Krupa, K. & Bekiesińska-Figatowska, M. Artifacts in magnetic resonance imaging. *Polish journal of radiology* **80**, 93 (2015).
29. Iglesias, J. E., Billot, B., Balbastre, Y., Magdamo, C., Arnold, S. E., Das, S., Edlow, B. L., Alexander, D. C., Golland, P. & Fischl, B. SynthSR: A public AI tool to turn heterogeneous clinical brain scans into high-resolution T1-weighted images for 3D morphometry. *Science advances* **9**, eadd3607 (2023).
30. Fischl, B. FreeSurfer. *Neuroimage* **62**, 774–781 (2012).
31. Smith, S. M., Vidaurre, D., Alfaro-Almagro, F., Nichols, T. E. & Miller, K. L. Estimation of brain age delta from brain imaging. *Neuroimage* **200**, 528–539 (2019).
32. Rosas, H. D., Lee, S. Y., Bender, A. C., Zaleta, A. K., Vangel, M., Yu, P., Fischl, B., Pappu, V., Onorato, C., Cha, J.-H., *et al.* Altered white matter microstructure in the corpus callosum in Huntington’s disease: implications for cortical “disconnection”. *Neuroimage* **49**, 2995–3004 (2010).
33. McCullagh, P. *Generalized linear models* (Routledge, 2019).
34. Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. & Baker, C. I. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience* **12**, 535–540 (2009).
35. Bryk, A. S. & Raudenbush, S. W. *Hierarchical linear models: applications and data analysis methods*. (Sage Publications, Inc, 1992).

36. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. *Densely connected convolutional networks* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), 4700–4708.
37. Peng, H., Gong, W., Beckmann, C. F., Vedaldi, A. & Smith, S. M. Accurate brain age prediction with lightweight deep neural networks. *Medical image analysis* **68**, 101871 (2021).
38. Ronneberger, O., Fischer, P. & Brox, T. *U-net: Convolutional networks for biomedical image segmentation* in *International Conference on Medical image computing and computer-assisted intervention* (2015), 234–241.
39. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020).
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **30** (2017).
41. Beaton, A. E. & Tukey, J. W. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics* **16**, 147–185 (1974).
42. Huber, P. J. in *Breakthroughs in statistics: Methodology and distribution* 492–518 (Springer, 1992).
43. Zhang, J., Yao, Y. & Deng, B. Fast and robust iterative closest point. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**, 3450–3466 (2021).
44. Besl, P. J. & McKay, N. D. *Method for registration of 3-D shapes in Sensor fusion IV: control paradigms and data structures* **1611** (1992), 586–606.
45. Mathieu, T. Concentration study of M-estimators using the influence function. *Electronic Journal of Statistics* **16**, 3695–3750 (2022).
46. Dice, L. R. Measures of the amount of ecologic association between species. *Ecology* **26**, 297–302 (1945).
47. Sorensen, T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biologiske skrifter* **5**, 1–34 (1948).
48. Birsan, T. & Tiba, D. *One hundred years since the introduction of the set distance by Dimitrie Pompeiu* in *IFIP Conference on System Modeling and Optimization* (2005), 35–39.
49. Jenkinson, M. Measuring transformation error by RMS deviation. *Studholme, C., Hill, DLG, Hawkes, DJ* (1999).
50. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**, 600–612 (2004).

51. Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics bulletin* **1**, 80–83 (1945).
52. Fisher, R. A. On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the royal statistical society* **85**, 87–94 (1922).
53. McHugh, M. L. The chi-square test of independence. *Biochemia medica* **23**, 143–149 (2013).
54. Bartko, J. J. The intraclass correlation coefficient as a measure of reliability. *Psychological reports* **19**, 3–11 (1966).
55. McGraw, K. O. & Wong, S. P. Forming inferences about some intraclass correlation coefficients. *Psychological methods* **1**, 30 (1996).
56. Reuter, M., Tisdall, M. D., Qureshi, A., Buckner, R. L., van der Kouwe, A. J. & Fischl, B. Head motion during MRI acquisition reduces gray matter volume and thickness estimates. *Neuroimage* **107**, 107–115 (2015).
57. Biller, A., Reuter, M., Patenaude, B., Homola, G., Breuer, F., Bendszus, M. & Bartsch, A. Responses of the human brain to mild dehydration and rehydration explored in vivo by 1H-MR imaging and spectroscopy. *American Journal of Neuro-radiology* **36**, 2277–2284 (2015).
58. Nakamura, K., Brown, R. A., Araujo, D., Narayanan, S. & Arnold, D. L. Correlation between brain volume change and T2 relaxation time induced by dehydration and rehydration: implications for monitoring atrophy in clinical studies. *NeuroImage: Clinical* **6**, 166–170 (2014).
59. Streitbürger, D.-P., Möller, H. E., Tittgemeyer, M., Hund-Georgiadis, M., Schroeter, M. L. & Mueller, K. Investigating structural brain changes of dehydration using voxel-based morphometry (2012).
60. Pardoe, H. R., Hiess, R. K. & Kuzniecky, R. Motion and morphometry in clinical and nonclinical populations. *Neuroimage* **135**, 177–185 (2016).
61. Alexander-Bloch, A., Clasen, L., Stockman, M., Ronan, L., Lalonde, F., Giedd, J. & Raznahan, A. Subtle in-scanner motion biases automated measurement of brain anatomy from in vivo MRI. *Human brain mapping* **37**, 2385–2397 (2016).
62. Savalia, N. K., Agres, P. F., Chan, M. Y., Feczko, E. J., Kennedy, K. M. & Wig, G. S. Motion-related artifacts in structural brain images revealed with independent estimates of in-scanner head motion. *Human brain mapping* **38**, 472–492 (2017).
63. Rosen, A. F., Roalf, D. R., Ruparel, K., Blake, J., Seelaus, K., Villa, L. P., Ciric, R., Cook, P. A., Davatzikos, C., Elliott, M. A., *et al.* Quantitative assessment of structural image quality. *Neuroimage* **169**, 407–418 (2018).

64. Gilmore, A. D., Buser, N. J. & Hanson, J. L. Variations in structural MRI quality significantly impact commonly used measures of brain anatomy. *Brain informatics* **8**, 1–15 (2021).
65. Stucht, D., Danishad, K. A., Schulze, P., Godenschweger, F., Zaitsev, M. & Speck, O. Highest resolution in vivo human brain MRI using prospective motion correction. *PloS one* **10**, e0133921 (2015).
66. Madan, C. R. Age differences in head motion and estimates of cortical morphology. *PeerJ* **6**, e5176 (2018).
67. Beyer, F., Prehn, K., Wüsten, K. A., Villringer, A., Ordemann, J., Flöel, A. & Witte, A. V. Weight loss reduces head motion: Revisiting a major confound in neuroimaging. *Human brain mapping* **41**, 2490–2494 (2020).
68. Beyer, F., Kharabian Masouleh, S., Huntenburg, J. M., Lampe, L., Luck, T., Riedel-Heller, S. G., Loeffler, M., Schroeter, M. L., Stumvoll, M., Villringer, A., *et al.* Higher body mass index is associated with reduced posterior default mode connectivity in older adults. *Human brain mapping* **38**, 3502–3515 (2017).
69. Cox, A. D., Virues-Ortega, J., Julio, F. & Martin, T. L. Establishing motion control in children with autism and intellectual disability: Applications for anatomical and functional MRI. *Journal of applied behavior analysis* **50**, 8–26 (2017).
70. Torres, E. B. & Denisova, K. Motor noise is rich signal in autism research and pharmacological treatments. *Scientific reports* **6**, 1–19 (2016).
71. Dosenbach, N. U., Koller, J. M., Earl, E. A., Miranda-Dominguez, O., Klein, R. L., Van, A. N., Snyder, A. Z., Nagel, B. J., Nigg, J. T., Nguyen, A. L., *et al.* Real-time motion analytics during brain MRI improve data quality and reduce costs. *Neuroimage* **161**, 80–93 (2017).
72. Thenganatt, M. A. & Jankovic, J. The relationship between essential tremor and Parkinson’s disease. *Parkinsonism & related disorders* **22**, S162–S165 (2016).
73. Versluis, M., Peeters, J., van Rooden, S., van der Grond, J., van Buchem, M. A., Webb, A. G. & van Osch, M. J. Origin and reduction of motion and f0 artifacts in high resolution T2*-weighted magnetic resonance imaging: application in Alzheimer’s disease patients. *Neuroimage* **51**, 1082–1088 (2010).
74. Rizk-Jackson, A., Stoffers, D., Sheldon, S., Kuperman, J., Dale, A., Goldstein, J., Corey-Bloom, J., Poldrack, R. A. & Aron, A. R. Evaluating imaging biomarkers for neurodegeneration in pre-symptomatic Huntington’s disease using machine learning techniques. *Neuroimage* **56**, 788–796 (2011).
75. Slipsager, J. M., Glimberg, S. L., Højgaard, L., Paulsen, R. R., Wighton, P., Tisdall, M. D., Jaimes, C., Gagoski, B. A., Grant, P. E., van Der Kouwe, A., *et al.* Comparison of prospective and retrospective motion correction in 3D-encoded neuroanatomical MRI. *Magnetic Resonance in Medicine* **87**, 629–645 (2022).

76. Breteler, M. M., Stöcker, T., Pracht, E., Brenner, D. & Stirnberg, R. MRI in the Rhineland study: a novel protocol for population neuroimaging. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* **10**, P92 (2014).
77. Stöcker, T. *Big data: the Rhineland study* in *Proceedings of the 24th Scientific Meeting of the International Society for Magnetic Resonance in Medicine (Singapore)* (2016).
78. Koch, A., Stirnberg, R., Estrada, S., Zeng, W., Lohner, V., Shahid, M., Ehses, P., Pracht, E., Reuter, M., Stoecker, T., *et al.* *Versatile MRI acquisition and processing protocol for population-based neuroimaging in Alzheimer's Association International Conference* (2023).
79. Olesen, O. V., Jørgensen, M. R., Paulsen, R. R., Højgaard, L., Roed, B. & Larsen, R. *Structured light 3D tracking system for measuring motions in PET brain imaging in Medical Imaging 2010: Visualization, Image-Guided Procedures, and Modeling* **7625** (2010), 286–296.
80. Olesen, O. V., Sullivan, J. M., Mulnix, T., Paulsen, R. R., Højgaard, L., Roed, B., Carson, R. E., Morris, E. D. & Larsen, R. List-mode PET motion correction using markerless head tracking: proof-of-concept with scans of human subject. *IEEE transactions on medical imaging* **32**, 200–209 (2012).
81. Arun, K. S., Huang, T. S. & Blostein, S. D. Least-squares fitting of two 3-D point sets. *IEEE Transactions on pattern analysis and machine intelligence*, 698–700 (1987).
82. Bergström, P. & Edlund, O. Robust registration of point sets using iteratively reweighted least squares. *Computational Optimization and Applications* **58**, 543–561 (2014).
83. Gramkow, C. On averaging rotations. *Journal of Mathematical Imaging and Vision* **15**, 7–16 (2001).
84. Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* **17**, 825–841 (2002).
85. Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A. & Gorgolewski, K. J. MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PloS one* **12**, e0184661 (2017).
86. Largent, A., Kapse, K., Barnett, S. D., *et al.* Image quality assessment of fetal brain MRI using multi-instance deep learning methods. *JMRI* **54**, 818–829 (2021).
87. Fantini, I., Rittner, L., Yasuda, C. & Lotufo, R. *Automatic detection of motion artifacts on MRI using Deep CNN* in *PRNI* (2018), 1–4.
88. Küstner, T., Jandt, M., Liebgott, A., Mauch, L., Martirosian, P., Bamberg, F., Nikolaou, K., Gatidis, S., Schick, F. & Yang, B. *Automatic motion artifact detection for Whole-Body magnetic resonance imaging* in *ICASSP* (2018), 995–999.

89. Sujit, S. J., Coronado, I., Kamali, A., Narayana, P. A. & Gabr, R. E. Automated image quality evaluation of structural brain MRI using an ensemble of deep learning networks. *Journal of Magnetic Resonance Imaging* **50**, 1260–1267 (2019).
90. Ma, J. J., Nakarmi, U., Kin, C. Y. S., Sandino, C. M., Cheng, J. Y., Syed, A. B., Wei, P., Pauly, J. M. & Vasanawala, S. S. *Diagnostic image quality assessment and classification in medical imaging: Opportunities and challenges in ISBI* (2020), 337–340.
91. Küstner, T., Liebgott, A., Mauch, L., Martirosian, P., Bamberg, F., Nikolaou, K., Yang, B., Schick, F. & Gatidis, S. Automated reference-free detection of motion artifacts in magnetic resonance images. *MAGMA* **31**, 243–256 (2018).
92. Stpień, I., Obuchowicz, R., Piórkowski, A. & Oszust, M. Fusion of Deep Convolutional Neural Networks for No-Reference Magnetic Resonance Image Quality Assessment. *Sensors* **21**, 1043 (2021).
93. Zukić, D., Haley, A., Lisle, C., Klo, J., Pohl, K. M., Johnson, H. J. & Chaudhary, A. *Medical Image Quality Assurance using Deep Learning in MIDL* (2022).
94. Lei, K., Syed, A. B., Zhu, X., Pauly, J. M. & Vasanawala, S. S. Artifact-and content-specific quality assessment for MRI with image rulers. *Medical Image Analysis* **77**, 102344 (2022).
95. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. *Densely connected convolutional networks* in CVPR (2017), 4700–4708.
96. Zaca, D., Hasson, U., Minati, L. & Jovicich, J. Method for retrospective estimation of natural head movement during structural MRI. *Journal of Magnetic Resonance Imaging* **48**, 927–937 (2018).
97. Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W. & Smith, S. M. Fsl. *Neuroimage* **62**, 782–790 (2012).
98. Ashburner, J. Computational anatomy with the SPM software. *Magnetic resonance imaging* **27**, 1163–1174 (2009).
99. Radwan, A. M., Emsell, L., Blommaert, J., Zhylka, A., Kovacs, S., Theys, T., Sollmann, N., Dupont, P. & Sunaert, S. Virtual brain grafting: Enabling whole brain parcellation in the presence of large lesions. *NeuroImage* **229**, 117731 (2021).
100. Guo, C., Ferreira, D., Fink, K., Westman, E. & Granberg, T. Repeatability and reproducibility of FreeSurfer, FSL-SIENAX and SPM brain volumetric measurements and the effect of lesion filling in multiple sclerosis. *European radiology* **29**, 1355–1364 (2019).
101. Subramanian, S., Gholami, A. & Biros, G. Simulation of glioblastoma growth using a 3D multispecies tumor model with mass effect. *Journal of mathematical biology* **79**, 941–967 (2019).

102. Zhang, S., Sun, H., Yang, X., Wan, X., Tan, Q., Li, S., Shao, H., Su, X., Yue, Q. & Gong, Q. An MRI Study Combining Virtual Brain Grafting and Surface-Based Morphometry Analysis to Investigate Contralateral Alterations in Cortical Morphology in Patients With Diffuse Low-Grade Glioma. *Journal of Magnetic Resonance Imaging* (2022).
103. Imms, P., Clemente, A., Deutscher, E., Radwan, A. M., Akhlaghi, H., Beech, P., Wilson, P. H., Irimia, A., Poudel, G., Domínguez Duque, J. F. & Caeyenberghs, K. Exploring personalized structural connectomics for moderate to severe traumatic brain injury. *Network Neuroscience* **7**, 160–183 (Jan. 2023).
104. Estrada, S., Kügler, D., Bahrami, E., Xu, P., Mousa, D., Breteler, M. M., Aziz, N. A. & Reuter, M. FastSurfer-HypVINN: Automated sub-segmentation of the hypothalamus and adjacent structures on high-resolution brain MRI. *Imaging Neuroscience* **1**, 1–32 (2023).
105. Faber, J., Kügler, D., Bahrami, E., Heinz, L.-S., Timmann, D., Ernst, T. M., Deike-Hofmann, K., Klockgether, T., van de Warrenburg, B., van Gaalen, J., Reetz, K., Romanzetti, S., Oz, G., Joers, J. M., Diedrichsen, J., Giunti, P., Garcia-Moreno, H., Jacobi, H., Jende, J., de Vries, J., Povazan, M., Barker, P. B., Steiner, K. M., Krahe, J. & Reuter, M. CerebNet: A fast and reliable deep-learning pipeline for detailed cerebellum sub-segmentation. *NeuroImage* **264**, 119703 (2022).
106. Baheti, B., Waldmannstetter, D., Chakrabarty, S., Akbari, M., Bilello, M., Wiestler, B., Schwarting, J., Calabrese, E., Jeffrey, R., Abidi, S., Mousa, M., Villanueva-Meyer, J., Marcus, D. S., Davathikos, C., Sotiras, A., Menze, B. H. & Bakas, S. *The Brain Tumor Sequence Registration Challenge: Establishing Correspondence between Pre-Operative and Follow-up MRI scans of diffuse glioma patients* technical report 2112.06979 (Dec. 2021).
107. Unterberger, I., Bauer, R., Walser, G. & Bauer, G. Corpus callosum and epilepsies. *Seizure* **37**, 55–60 (2016).
108. Piven, J., Bailey, J., Ranson, B. J., Arndt, S., *et al.* An MRI study of the corpus callosum in autism. *American Journal of Psychiatry* **154**, 1051–1056 (1997).
109. Woodruff, P., McManus, I. & David, A. Meta-analysis of corpus callosum size in schizophrenia. *Journal of Neurology, Neurosurgery & Psychiatry* **58**, 457–461 (1995).
110. Ozturk, A., Smith, S., Gordon-Lipkin, E., Harrison, D., Shiee, N., Pham, D., Caffo, B., Calabresi, P. & Reich, D. MRI of the corpus callosum in multiple sclerosis: association with disability. *Multiple Sclerosis Journal* **16**, 166–177 (2010).
111. Jaatela, J., Nurmi, T., Vallinoja, J., Mäenpää, H., Sairanen, V. & Piitulainen, H. Altered corpus callosum structure in adolescents with cerebral palsy: connection to gait and balance. *Brain Structure and Function* **228**, 1901–1915 (2023).

112. Yang, K., Wu, Z., Long, J., Li, W., Wang, X., Hu, N., Zhao, X. & Sun, T. White matter changes in Parkinson's disease. *npj Parkinson's Disease* **9**, 150 (2023).
113. Grande, I., Berk, M., Birmaher, B. & Vieta, E. Bipolar disorder. *The Lancet* **387**, 1561–1572 (2016).
114. Di Paola, M., Spalletta, G. & Caltagirone, C. In vivo structural neuroanatomy of corpus callosum in Alzheimer's disease and mild cognitive impairment using different MRI techniques: a review. *Journal of Alzheimer's disease* **20**, 67–95 (2010).
115. Bragin, A., Tcheng, T., Almajano, J. & Engel Jr, J. *Corpus callosum as a potential target for deep brain stimulation in EPILEPSIA* **48** (2007), 277–277.
116. Caverzasi, E., Papinutto, N., Cordano, C., Kirkish, G., Gundel, T. J., Zhu, A., Akula, A. V., Boscardin, W. J., Neeb, H., Henry, R. G., *et al.* MWF of the corpus callosum is a robust measure of remyelination: Results from the ReBUILD trial. *Proceedings of the National Academy of Sciences* **120**, e2217635120 (2023).
117. Adamson, C., Beare, R., Walterfang, M. & Seal, M. Software pipeline for mid-sagittal corpus callosum thickness profile processing: automated segmentation, manual editor, thickness profile generator, group-wise statistical comparison and results display. *Neuroinformatics* **12**, 595–614 (2014).
118. Ardekani, B., Lee, S. & Bachman, A. *Automated segmentation of corpus callosum in MRI scans: application to tracking alzheimer's disease progression in Conference Abstract: ACNS-2013 Australasian Cognitive Neuroscience Society Conference* (2013).
119. Baudo, M. C., Colombo, E. E., Pérez, M. A., Rahn, S. I., Domenech, N. C., Chuang, J., Forlizzi, V., Buosi, P., Baldoncini, M. & Bendersky, M. Three-dimensional anatomy of the anterior commissure: a Tractography and anatomical study. *World neurosurgery* **159**, e365–e374 (2022).
120. Jäncke, L., Staiger, J. F., Schlaug, G., Huang, Y. & Steinmetz, H. The relationship between corpus callosum size and forebrain volume. *Cerebral cortex (New York, NY: 1991)* **7**, 48–56 (1997).
121. Vermeulen, C. L., du Toit, P. J., Venter, G. & Human-Baron, R. A morphological study of the shape of the corpus callosum in normal, schizophrenic and bipolar patients. *Journal of anatomy* **242**, 153–163 (2022).
122. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely Connected Convolutional Networks, Aug. *arXiv preprint arXiv:1608.06993* (2016).
123. Ardekani, B. A., Kershaw, J., Braun, M. & Kanuo, I. Automatic detection of the mid-sagittal plane in 3-D brain images. *IEEE transactions on medical imaging* **16**, 947–952 (1997).

124. Yaldizli, Ö., Atefy, R., Gass, A., Sturm, D., Glassl, S., Tettenborn, B. & Putzki, N. Corpus callosum index and long-term disability in multiple sclerosis patients. *Journal of neurology* **257**, 1256–1264 (2010).
125. Van Schependom, J., Niemantsverdriet, E., Smeets, D. & Engelborghs, S. Callosal circularity as an early marker for Alzheimer’s disease. *NeuroImage: Clinical* **19**, 516–526 (2018).
126. Witelson, S. F. The brain connection: the corpus callosum is larger in left-handers. *Science* **229**, 665–668 (1985).
127. Hofer, S. & Frahm, J. Topography of the human corpus callosum revisited—comprehensive fiber tractography using diffusion tensor magnetic resonance imaging. *Neuroimage* **32**, 989–994 (2006).
128. Hampel, H., Teipel, S. J., Alexander, G. E., Horwitz, B., Teichberg, D., Schapiro, M. B. & Rapoport, S. I. Corpus callosum atrophy is a possible indicator of region- and cell type-specific neuronal degeneration in Alzheimer disease: A magnetic resonance imaging analysis. *Archives of neurology* **55**, 193–198 (1998).
129. Greve, D. N., Billot, B., Cordero, D., Hoopes, A., Hoffmann, M., Dalca, A. V., Fischl, B., Iglesias, J. E. & Augustinack, J. C. A deep learning toolbox for automatic segmentation of subcortical limbic structures from MRI images. *Neuroimage* **244**, 118610 (2021).
130. Chang, J., Steyn, F., Ngo, S., Henderson, R., Guo, C., Bollmann, S., Fripp, J., Barth, M. & Shaw, T. Open-source hypothalamic-Fornix (OSHy-X) atlases and segmentation tool for 3T and 7T. *Journal of Open Source Software* **7**, 4368 (2022).
131. Bricout, C., Kahou, S. E. & Bouix, S. *Estimation of Head Motion in Structural MRI and its Impact on Cortical Thickness Measurements in Retrospective Data* 2025. arXiv: [2505.23916](https://arxiv.org/abs/2505.23916) [eess.IV].
132. Wolleb, J., Bieder, F., Sandkühler, R. & Cattin, P. C. *Diffusion models for medical anomaly detection in International Conference on Medical image computing and computer-assisted intervention* (2022), 35–45.

Bibliography for the publication: Quantifying MR head motion in the Rhineland Study – A robust method for population cohorts

1. Henschel, L., Kügler, D. & Reuter, M. FastSurferVINN: Building resolution-independence into deep learning segmentation methods—A solution for HighRes brain MRI. *NeuroImage* **251**, 118933 (2022).
2. Fischl, B. FreeSurfer. *Neuroimage* **62**, 774–781 (2012).
3. Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W. & Smith, S. M. Fsl. *Neuroimage* **62**, 782–790 (2012).
4. Reuter, M., Schmansky, N. J., Rosas, H. D. & Fischl, B. Within-subject template estimation for unbiased longitudinal image analysis. *Neuroimage* **61**, 1402–1418 (2012).
5. Biller, A., Reuter, M., Patenaude, B., Homola, G., Breuer, F., Bendszus, M. & Bartsch, A. Responses of the human brain to mild dehydration and rehydration explored in vivo by 1H-MR imaging and spectroscopy. *American Journal of Neuro-radiology* **36**, 2277–2284 (2015).
6. Nakamura, K., Brown, R. A., Araujo, D., Narayanan, S. & Arnold, D. L. Correlation between brain volume change and T2 relaxation time induced by dehydration and rehydration: implications for monitoring atrophy in clinical studies. *NeuroImage: Clinical* **6**, 166–170 (2014).
7. Streitbürger, D.-P., Möller, H. E., Tittgemeyer, M., Hund-Georgiadis, M., Schroeter, M. L. & Mueller, K. Investigating structural brain changes of dehydration using voxel-based morphometry (2012).
8. Reuter, M., Tisdall, M. D., Qureshi, A., Buckner, R. L., van der Kouwe, A. J. & Fischl, B. Head motion during MRI acquisition reduces gray matter volume and thickness estimates. *Neuroimage* **107**, 107–115 (2015).
9. Pardoe, H. R., Hiess, R. K. & Kuzniecky, R. Motion and morphometry in clinical and nonclinical populations. *Neuroimage* **135**, 177–185 (2016).
10. Alexander-Bloch, A., Clasen, L., Stockman, M., Ronan, L., Lalonde, F., Giedd, J. & Raznahan, A. Subtle in-scanner motion biases automated measurement of brain anatomy from in vivo MRI. *Human brain mapping* **37**, 2385–2397 (2016).

11. Savalia, N. K., Agres, P. F., Chan, M. Y., Feczko, E. J., Kennedy, K. M. & Wig, G. S. Motion-related artifacts in structural brain images revealed with independent estimates of in-scanner head motion. *Human brain mapping* **38**, 472–492 (2017).
12. Rosen, A. F., Roalf, D. R., Ruparel, K., Blake, J., Seelaus, K., Villa, L. P., Ciric, R., Cook, P. A., Davatzikos, C., Elliott, M. A., *et al.* Quantitative assessment of structural image quality. *Neuroimage* **169**, 407–418 (2018).
13. Gilmore, A. D., Buser, N. J. & Hanson, J. L. Variations in structural MRI quality significantly impact commonly used measures of brain anatomy. *Brain informatics* **8**, 1–15 (2021).
14. Stucht, D., Danishad, K. A., Schulze, P., Godenschweger, F., Zaitsev, M. & Speck, O. Highest resolution in vivo human brain MRI using prospective motion correction. *PloS one* **10**, e0133921 (2015).
15. Madan, C. R. Age differences in head motion and estimates of cortical morphology. *PeerJ* **6**, e5176 (2018).
16. Beyer, F., Prehn, K., Wüsten, K. A., Villringer, A., Ordemann, J., Flöel, A. & Witte, A. V. Weight loss reduces head motion: Revisiting a major confound in neuroimaging. *Human brain mapping* **41**, 2490–2494 (2020).
17. Beyer, F., Kharabian Masouleh, S., Huntenburg, J. M., Lampe, L., Luck, T., Riedel-Heller, S. G., Loeffler, M., Schroeter, M. L., Stumvoll, M., Villringer, A., *et al.* Higher body mass index is associated with reduced posterior default mode connectivity in older adults. *Human brain mapping* **38**, 3502–3515 (2017).
18. Cox, A. D., Virues-Ortega, J., Julio, F. & Martin, T. L. Establishing motion control in children with autism and intellectual disability: Applications for anatomical and functional MRI. *Journal of applied behavior analysis* **50**, 8–26 (2017).
19. Torres, E. B. & Denisova, K. Motor noise is rich signal in autism research and pharmacological treatments. *Scientific reports* **6**, 1–19 (2016).
20. Dosenbach, N. U., Koller, J. M., Earl, E. A., Miranda-Dominguez, O., Klein, R. L., Van, A. N., Snyder, A. Z., Nagel, B. J., Nigg, J. T., Nguyen, A. L., *et al.* Real-time motion analytics during brain MRI improve data quality and reduce costs. *Neuroimage* **161**, 80–93 (2017).
21. Thenganatt, M. A. & Jankovic, J. The relationship between essential tremor and Parkinson's disease. *Parkinsonism & related disorders* **22**, S162–S165 (2016).

22. Versluis, M., Peeters, J., van Rooden, S., van der Grond, J., van Buchem, M. A., Webb, A. G. & van Osch, M. J. Origin and reduction of motion and f0 artifacts in high resolution T2*-weighted magnetic resonance imaging: application in Alzheimer's disease patients. *Neuroimage* **51**, 1082–1088 (2010).
23. Rizk-Jackson, A., Stoffers, D., Sheldon, S., Kuperman, J., Dale, A., Goldstein, J., Corey-Bloom, J., Poldrack, R. A. & Aron, A. R. Evaluating imaging biomarkers for neurodegeneration in pre-symptomatic Huntington's disease using machine learning techniques. *Neuroimage* **56**, 788–796 (2011).
24. Slipsager, J. M., Ellegaard, A. H., Glimberg, S. L., Paulsen, R. R., Tisdall, M. D., Wighton, P., van der Kouwe, A., Marner, L., Henriksen, O. M., Law, I., *et al.* Markerless motion tracking and correction for PET, MRI, and simultaneous PET/MRI. *Plos one* **14**, e0215524 (2019).
25. Slipsager, J. M., Glimberg, S. L., Højgaard, L., Paulsen, R. R., Wighton, P., Tisdall, M. D., Jaimes, C., Gagoski, B. A., Grant, P. E., van Der Kouwe, A., *et al.* Comparison of prospective and retrospective motion correction in 3D-encoded neuroanatomical MRI. *Magnetic Resonance in Medicine* **87**, 629–645 (2022).
26. Breteler, M. M., Stöcker, T., Pracht, E., Brenner, D. & Stirnberg, R. MRI in the Rhineland study: a novel protocol for population neuroimaging. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* **10**, P92 (2014).
27. Stöcker, T. *Big data: the Rhineland study* in *Proceedings of the 24th Scientific Meeting of the International Society for Magnetic Resonance in Medicine (Singapore)* (2016).
28. Pollak, C., Kügler, D. & Reuter, M. *Estimating Head Motion from Mr-Images* in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)* (2023), 1–5.
29. Zaitsev, M., Maclaren, J. & Herbst, M. Motion artifacts in MRI: A complex problem with many partial solutions. *Journal of magnetic resonance imaging : JMRI* **42**, 887–901 (2015).
30. Dbouk, T. & Drikakis, D. On coughing and airborne droplet transmission to humans. *Physics of Fluids* **32**, 053310 (2020).
31. Eberhardt, O. & Topka, H. Myoclonic disorders. *Brain Sciences* **7**, 103 (2017).
32. Maknojia, S., Churchill, N. W., Schweizer, T. A. & Graham, S. Resting state fMRI: Going through the motions. *Frontiers in neuroscience* **13**, 825 (2019).
33. Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* **17**, 825–841 (2002).

34. Goto, M., Abe, O., Miyati, T., Yamasue, H., Gomi, T. & Takeda, T. Head motion and correction methods in resting-state functional MRI. *Magnetic Resonance in Medical Sciences*, rev-2015 (2015).
35. Andronesi, O. C., Bhattacharyya, P. K., Bogner, W., Choi, I.-Y., Hess, A. T., Lee, P., Meintjes, E. M., Tisdall, M. D., Zaitzev, M. & van der Kouwe, A. Motion correction methods for MRS: experts' consensus recommendations. *NMR in Biomedicine* **34**, e4364 (2021).
36. Kober, T., Marques, J. P., Gruetter, R. & Krueger, G. Head motion detection using FID navigators. *Magnetic resonance in medicine* **66**, 135–143 (2011).
37. Fu, Z. W., Wang, Y., Grimm, R. C., Rossman, P. J., Felmlee, J. P., Riederer, S. J. & Ehman, R. L. Orbital navigator echoes for motion measurements in magnetic resonance imaging. *Magnetic resonance in medicine* **34**, 746–753 (1995).
38. Costa, A. F., Petrie, D. W., Yen, Y.-F. & Drangova, M. Using the axis of rotation of polar navigator echoes to rapidly measure 3D rigid-body motion. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* **53**, 150–158 (2005).
39. Welch, E. B., Manduca, A., Grimm, R. C., Ward, H. A. & Jack Jr, C. R. Spherical navigator echoes for full 3D rigid body motion measurement in MRI. *Magnetic resonance in medicine: an official journal of the international society for magnetic resonance in medicine* **47**, 32–41 (2002).
40. Van der Kouwe, A. J., Benner, T. & Dale, A. M. Real-time rigid body motion correction and shimming using cloverleaf navigators. *Magnetic resonance in medicine: an official journal of the international society for magnetic resonance in medicine* **56**, 1019–1032 (2006).
41. Tisdall, M. D., Hess, A. T., Reuter, M., Meintjes, E. M., Fischl, B. & van der Kouwe, A. J. Volumetric navigators for prospective motion correction and selective reacquisition in neuroanatomical MRI. *Magnetic resonance in medicine* **68**, 389–399 (2012).
42. Hess, A. T., Dylan Tisdall, M., Andronesi, O. C., Meintjes, E. M. & van der Kouwe, A. J. Real-time motion and B0 corrected single voxel spectroscopy using volumetric navigators. *Magnetic resonance in medicine* **66**, 314–323 (2011).
43. Alhamud, A., Tisdall, M. D., Hess, A. T., Hasan, K. M., Meintjes, E. M. & van der Kouwe, A. J. Volumetric navigators for real-time motion correction in diffusion tensor imaging. *Magnetic resonance in medicine* **68**, 1097–1108 (2012).

44. Forman, C., Aksoy, M., Horneegger, J. & Bammer, R. Self-encoded marker for optical prospective head motion correction in MRI. *Medical image analysis* **15**, 708–719 (2011).
45. Maclaren, J., Armstrong, B. S., Barrows, R. T., Danishad, K., Ernst, T., Foster, C. L., Gumus, K., Herbst, M., Kadashevich, I. Y., Kusik, T. P., *et al.* Measurement and correction of microscopic head motion during magnetic resonance imaging of the brain. *PloS one* **7**, e48088 (2012).
46. Todd, N., Josephs, O., Callaghan, M. F., Lutti, A. & Weiskopf, N. Prospective motion correction of 3D echo-planar imaging data for functional MRI using optical tracking. *NeuroImage* **113**, 1–12 (2015).
47. Aksoy, M., Maclaren, J. & Bammer, R. Prospective motion correction for 3D pseudo-continuous arterial spin labeling using an external optical tracking system. *Magnetic resonance imaging* **39**, 44–52 (2017).
48. Herbst, M., Lovell-Smith, C., Haeublein, B., Sostheim, R., Maclaren, J. R., Korvink, J. G. & Zaitsev, M. *On the robustness of prospective motion correction for clinical routine* in *Proceedings of the 21st Scientific Meeting: International Society for Magnetic Resonance in Medicine* (2013), 3766.
49. Olesen, O. V., Jørgensen, M. R., Paulsen, R. R., Højgaard, L., Roed, B. & Larsen, R. *Structured light 3D tracking system for measuring motions in PET brain imaging* in *Medical Imaging 2010: Visualization, Image-Guided Procedures, and Modeling* **7625** (2010), 286–296.
50. Pardoe, H. R., Martin, S. P., Zhao, Y., George, A., Yuan, H., Zhou, J., Liu, W. & Devinsky, O. Estimation of in-scanner head pose changes during structural MRI using a convolutional neural network trained on eye tracker video. *Magnetic Resonance Imaging* **81**, 101–108 (2021).
51. Kyme, A. Z., Aksoy, M., Henry, D. L., Bammer, R. & Maclaren, J. Marker-free optical stereo motion tracking for in-bore MRI and PET-MRI application. *Medical physics* **47**, 3321–3331 (2020).
52. Laustsen, M., Andersen, M., Xue, R., Madsen, K. H. & Hanson, L. G. Tracking of rigid head motion during MRI using an EEG system. *Magnetic Resonance in Medicine* (2022).
53. Musa, M., Sengupta, S. & Chen, Y. MRI-Compatible Soft Robotic Sensing Pad for Head Motion Detection. *IEEE Robotics and Automation Letters* **7**, 3632–3639 (2022).

54. Einspänner, E., Jochimsen, T. H., Harries, J., Melzer, A., Unger, M., Brown, R., Thielemans, K., Sabri, O. & Sattler, B. Evaluating different methods of MR-based motion correction in simultaneous PET/MR using a head phantom moved by a robotic system. *EJNMMI physics* **9**, 1–17 (2022).
55. Hodgson, K., Poldrack, R. A., Curran, J. E., Knowles, E. E., Mathias, S., Göring, H. H., Yao, N., Olvera, R. L., Fox, P. T., Almasy, L., *et al.* Shared genetic factors influence head motion during MRI and body mass index. *Cerebral Cortex* **27**, 5539–5546 (2017).
56. Wylie, G. R., Genova, H., DeLuca, J., Chiaravalloti, N. & Sumowski, J. F. Functional magnetic resonance imaging movers and shakers: Does subject-movement cause sampling bias? *Human brain mapping* **35**, 1–13 (2014).
57. Couvy-Duchesne, B., Ebejer, J. L., Gillespie, N. A., Duffy, D. L., Hickie, I. B., Thompson, P. M., Martin, N. G., de Zubicaray, G. I., McMahon, K. L., Medland, S. E., *et al.* Head motion and inattention/hyperactivity share common genetic influences: implications for fMRI studies of ADHD. *PloS one* **11**, e0146271 (2016).
58. Kong, X.-z., Zhen, Z., Li, X., Lu, H.-h., Wang, R., Liu, L., He, Y., Zang, Y. & Liu, J. Individual differences in impulsivity predict head motion during magnetic resonance imaging. *PloS one* **9**, e104989 (2014).
59. Ekhtiari, H., Kuplicki, R., Yeh, H.-w. & Paulus, M. P. Physical characteristics not psychological state or trait characteristics predict motion during resting state fMRI. *Scientific reports* **9**, 1–8 (2019).
60. Afacan, O., Erem, B., Roby, D. P., Roth, N., Roth, A., Prabhu, S. P. & Warfield, S. K. Evaluation of motion and its effect on brain magnetic resonance image quality in children. *Pediatric radiology* **46**, 1728–1735 (2016).
61. Janos, S., Schooler, G. R., Ngo, J. S. & Davis, J. T. Free-breathing unsedated MRI in children: Justification and techniques. *Journal of Magnetic Resonance Imaging* **50**, 365–376 (2019).
62. Barkovich, M. J., Li, Y., Desikan, R. S., Barkovich, A. J. & Xu, D. Challenges in pediatric neuroimaging. *Neuroimage* **185**, 793–801 (2019).
63. Oztek, M. A., Noda, S., Beauchemin, E. A. & Otto, R. K. Gentle touch: noninvasive approaches to improve patient comfort and cooperation for pediatric imaging. *Topics in Magnetic Resonance Imaging* **29**, 187–195 (2020).
64. Chan, M. Y., Park, D. C., Savalia, N. K., Petersen, S. E. & Wig, G. S. Decreased segregation of brain systems across the healthy adult lifespan. *Proceedings of the National Academy of Sciences* **111**, E4997–E5006 (2014).

65. Andrews-Hanna, J. R., Snyder, A. Z., Vincent, J. L., Lustig, C., Head, D., Raichle, M. E. & Buckner, R. L. Disruption of large-scale brain systems in advanced aging. *Neuron* **56**, 924–935 (2007).
66. Couvy-Duchesne, B., Blokland, G. A., Hickie, I. B., Thompson, P. M., Martin, N. G., de Zubicaray, G. I., McMahon, K. L. & Wright, M. J. Heritability of head motion during resting state functional MRI in 462 healthy twins. *Neuroimage* **102**, 424–434 (2014).
67. Engelhardt, L. E., Roe, M. A., Juranek, J., DeMaster, D., Harden, K. P., Tucker-Drob, E. M. & Church, J. A. Children’s head motion during fMRI tasks is heritable and stable over time. *Developmental cognitive neuroscience* **25**, 58–68 (2017).
68. Zeng, L.-L., Wang, D., Fox, M. D., Sabuncu, M., Hu, D., Ge, M., Buckner, R. L. & Liu, H. Neurobiological basis of head motion in brain imaging. *Proceedings of the National Academy of Sciences* **111**, 6058–6062 (2014).
69. Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A. & Gorgolewski, K. J. MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PloS one* **12**, e0184661 (2017).
70. Pannetier, N. A., Stavrinos, T., Ng, P., Herbst, M., Zaitsev, M., Young, K., Matson, G. & Schuff, N. Quantitative framework for prospective motion correction evaluation. *Magnetic resonance in medicine* **75**, 810–816 (2016).
71. Callaghan, M. F., Josephs, O., Herbst, M., Zaitsev, M., Todd, N. & Weiskopf, N. An evaluation of prospective motion correction (PMC) for high resolution quantitative MRI. *Frontiers in neuroscience* **9**, 97 (2015).
72. Zaca, D., Hasson, U., Minati, L. & Jovicich, J. Method for retrospective estimation of natural head movement during structural MRI. *Journal of Magnetic Resonance Imaging* **48**, 927–937 (2018).
73. Zukić, D., Haley, A., Lisle, C., Klo, J., Pohl, K. M., Johnson, H. J. & Chaudhary, A. *Medical Image Quality Assurance using Deep Learning in MIDL* (2022).
74. Lei, K., Syed, A. B., Zhu, X., Pauly, J. M. & Vasanawala, S. S. Artifact-and content-specific quality assessment for MRI with image rulers. *Medical Image Analysis* **77**, 102344 (2022).
75. Küstner, T., Jandt, M., Liebgott, A., Mauch, L., Martirosian, P., Bamberg, F., Nikolaou, K., Gatidis, S., Schick, F. & Yang, B. *Automatic motion artifact detection for Whole-Body magnetic resonance imaging in ICASSP* (2018), 995–999.

76. Largent, A., Kapse, K., Barnett, S. D., *et al.* Image quality assessment of fetal brain MRI using multi-instance deep learning methods. *JMRI* **54**, 818–829 (2021).
77. Fantini, I., Rittner, L., Yasuda, C. & Lotufo, R. *Automatic detection of motion artifacts on MRI using Deep CNN* in *PRNI* (2018), 1–4.
78. Sujit, S. J., Coronado, I., Kamali, A., Narayana, P. A. & Gabr, R. E. Automated image quality evaluation of structural brain MRI using an ensemble of deep learning networks. *Journal of Magnetic Resonance Imaging* **50**, 1260–1267 (2019).
79. Ma, J. J., Nakarmi, U., Kin, C. Y. S., Sandino, C. M., Cheng, J. Y., Syed, A. B., Wei, P., Pauly, J. M. & Vasanawala, S. S. *Diagnostic image quality assessment and classification in medical imaging: Opportunities and challenges* in *ISBI* (2020), 337–340.
80. Küstner, T., Liebgott, A., Mauch, L., Martirosian, P., Bamberg, F., Nikolaou, K., Yang, B., Schick, F. & Gatidis, S. Automated reference-free detection of motion artifacts in magnetic resonance images. *MAGMA* **31**, 243–256 (2018).
81. Stpień, I., Obuchowicz, R., Piórkowski, A. & Oszust, M. Fusion of Deep Convolutional Neural Networks for No-Reference Magnetic Resonance Image Quality Assessment. *Sensors* **21**, 1043 (2021).
82. Sciarra, A., Chatterjee, S., Dünnwald, M., Placidi, G., Nürnberger, A., Speck, O. & Oeltze-Jafra, S. *Reference-less SSIM Regression for Detection and Quantification of Motion Artefacts in Brain MRIs* in *MIDL* (2022).
83. Castella, R., Arn, L., Dupuis, E., Callaghan, M. F., Draganski, B. & Lutti, A. Controlling motion artefact levels in MR images by suspending data acquisition during periods of head motion. *Magnetic resonance in medicine* **80**, 2415–2426 (2018).
84. Van der Kouwe, A. J., Benner, T., Salat, D. H. & Fischl, B. Brain morphometry with multiecho MPRAGE. *Neuroimage* **40**, 559–569 (2008).
85. Brenner, D., Stirnberg, R., Pracht, E. D. & Stöcker, T. Two-dimensional accelerated MP-RAGE imaging with flexible linear reordering. *Magnetic Resonance Materials in Physics, Biology and Medicine* **27**, 455–462 (2014).
86. Mugler III, J. P. Optimized three-dimensional fast-spin-echo MRI. *Journal of magnetic resonance imaging* **39**, 745–767 (2014).
87. Stirnberg, R., Huijbers, W., Brenner, D., Poser, B. A., Breteler, M. & Stöcker, T. Rapid whole-brain resting-state fMRI at 3 T: Efficiency-optimized three-dimensional EPI versus repetition time-matched simultaneous-multi-slice EPI. *Neuroimage* **163**, 81–92 (2017).

88. Olesen, O. V., Sullivan, J. M., Mulnix, T., Paulsen, R. R., Hojgaard, L., Roed, B., Carson, R. E., Morris, E. D. & Larsen, R. List-mode PET motion correction using markerless head tracking: proof-of-concept with scans of human subject. *IEEE transactions on medical imaging* **32**, 200–209 (2012).
89. Huijbers, W., Van Dijk, K. R., Boenniger, M. M., Stirnberg, R. & Breteler, M. M. Less head motion during MRI under task than resting-state conditions. *Neuroimage* **147**, 111–120 (2017).
90. Vanderwal, T., Kelly, C., Eilbott, J., Mayes, L. C. & Castellanos, F. X. Inscapes: A movie paradigm to improve compliance in functional magnetic resonance imaging. *Neuroimage* **122**, 222–232 (2015).
91. Greene, D. J., Koller, J. M., Hampton, J. M., Wesevich, V., Van, A. N., Nguyen, A. L., Hoyt, C. R., McIntyre, L., Earl, E. A., Klein, R. L., *et al.* Behavioral interventions for reducing head motion during MRI scans in children. *Neuroimage* **171**, 234–245 (2018).
92. Arun, K. S., Huang, T. S. & Blostein, S. D. Least-squares fitting of two 3-D point sets. *IEEE Transactions on pattern analysis and machine intelligence*, 698–700 (1987).
93. Bergström, P. & Edlund, O. Robust registration of point sets using iteratively reweighted least squares. *Computational Optimization and Applications* **58**, 543–561 (2014).
94. Frost, R., Wighton, P., Karahanoğlu, F. I., Robertson, R. L., Grant, P. E., Fischl, B., Tisdall, M. D. & van der Kouwe, A. Markerless high-frequency prospective motion correction for neuroanatomical MRI. *Magnetic resonance in medicine* **82**, 126–144 (2019).
95. Reuter, M., Rosas, H. D. & Fischl, B. Highly accurate inverse consistent registration: A robust approach. *NeuroImage* **53**, 1181–1196 (2010).
96. Greve, D. N. & Fischl, B. Accurate and robust brain image alignment using boundary-based registration. *Neuroimage* **48**, 63–72 (2009).
97. Gramkow, C. On averaging rotations. *Journal of Mathematical Imaging and Vision* **15**, 7–16 (2001).
98. Wilke, M. An alternative approach towards assessing and accounting for individual motion in fMRI timeseries. *Neuroimage* **59**, 2062–2072 (2012).
99. Wilke, M. Isolated assessment of translation or rotation severely underestimates the effects of subject motion in fMRI data. *PloS one* **9**, e106498 (2014).

100. Jenkinson, M. Measuring transformation error by RMS deviation. *Studholme, C., Hill, DLG, Hawkes, DJ* (1999).
101. Kraskov, A., Stögbauer, H. & Grassberger, P. Estimating mutual information. *Physical review E* **69**, 066138 (2004).
102. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
103. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods* **17**, 261–272 (2020).
104. Seabold, S. & Perktold, J. *statsmodels: Econometric and statistical modeling with python* in *9th Python in Science Conference* (2010).
105. Atkinson, D., Hill, D. L., Stoyale, P. N., Summers, P. E. & Keevil, S. F. Automatic correction of motion artifacts in magnetic resonance images using an entropy focus criterion. *IEEE Transactions on Medical imaging* **16**, 903–910 (1997).
106. Mortamet, B., Bernstein, M. A., Jack Jr, C. R., Gunter, J. L., Ward, C., Britson, P. J., Meuli, R., Thiran, J.-P. & Krueger, G. Automatic quality assessment in structural brain magnetic resonance imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* **62**, 365–372 (2009).
107. Shehzad, Z., Giavasis, S., Li, Q., Benhajali, Y., Yan, C., Yang, Z., Milham, M., Bellec, P. & Craddock, C. The Preprocessed Connectomes Project Quality Assessment Protocol—a resource for measuring the quality of MRI data. *Frontiers in neuroscience* **47** (2015).
108. Provins, C., Schöttner, M., Dayan, M., Nastase, V., Lunde, J., Benach, O. M., MacNicol, E., Savary, E., Seeley, S. H., Hagen, M. P., *et al.* Signal-to-noise ratio estimates predict head motion presence in T1-weighted MRI (2023).
109. Meissner, T. W., Walbrin, J., Nordt, M., Koldewyn, K. & Weigelt, S. Head motion during fMRI tasks is reduced in children and adults if participants take breaks. *Developmental cognitive neuroscience* **44**, 100803 (2020).
110. Littleton, S. W. Impact of obesity on respiratory function. *Respirology* **17**, 43–49 (2012).

111. Tomasi, D. & Volkow, N. D. Striatocortical pathway dysfunction in addiction and obesity: differences and similarities. *Critical reviews in biochemistry and molecular biology* **48**, 1–19 (2013).
112. Frew, S., Samara, A., Shearer, H., Eilbott, J. & Vanderwal, T. Getting the nod: Pediatric head motion in a transdiagnostic sample during movie-and resting-state fMRI. *PloS one* **17**, e0265112 (2022).

Bibliography for the publication: Estimating head motion from MR-images

1. Reuter, M., Tisdall, M. D., Qureshi, A., Buckner, R. L., van der Kouwe, A. J. & Fischl, B. Head motion during MRI acquisition reduces gray matter volume and thickness estimates. *Neuroimage* **107**, 107–115 (2015).
2. Alexander-Bloch, A., Clasen, L., Stockman, M., Ronan, L., Lalonde, F., Giedd, J. & Raznahan, A. Subtle in-scanner motion biases automated measurement of brain anatomy from in vivo MRI. *Human brain mapping* **37**, 2385–2397 (2016).
3. Gilmore, A. D., Buser, N. J. & Hanson, J. L. Variations in structural MRI quality significantly impact commonly used measures of brain anatomy. *Brain informatics* **8**, 1–15 (2021).
4. Savalia, N. K., Agres, P. F., Chan, M. Y., Feczko, E. J., Kennedy, K. M. & Wig, G. S. Motion-related artifacts in structural brain images revealed with independent estimates of in-scanner head motion. *Human brain mapping* **38**, 472–492 (2017).
5. Pardoe, H. R., Hiess, R. K. & Kuzniecky, R. Motion and morphometry in clinical and nonclinical populations. *Neuroimage* **135**, 177–185 (2016).
6. Rosen, A. F., Roalf, D. R., Ruparel, K., Blake, J., Seelaus, K., Villa, L. P., Ciric, R., Cook, P. A., Davatzikos, C., Elliott, M. A., *et al.* Quantitative assessment of structural image quality. *Neuroimage* **169**, 407–418 (2018).
7. Zaca, D., Hasson, U., Minati, L. & Jovicich, J. Method for retrospective estimation of natural head movement during structural MRI. *Journal of Magnetic Resonance Imaging* **48**, 927–937 (2018).
8. Zukić, D., Haley, A., Lisle, C., Klo, J., Pohl, K. M., Johnson, H. J. & Chaudhary, A. *Medical Image Quality Assurance using Deep Learning in MIDL* (2022).

9. Küstner, T., Jandt, M., Liebgott, A., Mauch, L., Martirosian, P., Bamberg, F., Nikolaou, K., Gatidis, S., Schick, F. & Yang, B. *Automatic motion artifact detection for Whole-Body magnetic resonance imaging in ICASSP* (2018), 995–999.
10. Fantini, I., Rittner, L., Yasuda, C. & Lotufo, R. *Automatic detection of motion artifacts on MRI using Deep CNN in PRNI* (2018), 1–4.
11. Breteler, M. M., Stöcker, T., Pracht, E., Brenner, D. & Stirnberg, R. MRI in the Rhineland study: a novel protocol for population neuroimaging. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* **10**, P92 (2014).
12. Stöcker, T. *Big data: the Rhineland study in Proceedings of the 24th Scientific Meeting of the International Society for Magnetic Resonance in Medicine (Singapore)* (2016).
13. Largent, A., Kapse, K., Barnett, S. D., *et al.* Image quality assessment of fetal brain MRI using multi-instance deep learning methods. *JMRI* **54**, 818–829 (2021).
14. Sujit, S. J., Coronado, I., Kamali, A., Narayana, P. A. & Gabr, R. E. Automated image quality evaluation of structural brain MRI using an ensemble of deep learning networks. *Journal of Magnetic Resonance Imaging* **50**, 1260–1267 (2019).
15. Ma, J. J., Nakarmi, U., Kin, C. Y. S., Sandino, C. M., Cheng, J. Y., Syed, A. B., Wei, P., Pauly, J. M. & Vasanawala, S. S. *Diagnostic image quality assessment and classification in medical imaging: Opportunities and challenges in ISBI* (2020), 337–340.
16. Küstner, T., Liebgott, A., Mauch, L., Martirosian, P., Bamberg, F., Nikolaou, K., Yang, B., Schick, F. & Gatidis, S. Automated reference-free detection of motion artifacts in magnetic resonance images. *MAGMA* **31**, 243–256 (2018).
17. Stpień, I., Obuchowicz, R., Piórkowski, A. & Oszust, M. Fusion of Deep Convolutional Neural Networks for No-Reference Magnetic Resonance Image Quality Assessment. *Sensors* **21**, 1043 (2021).
18. Lei, K., Syed, A. B., Zhu, X., Pauly, J. M. & Vasanawala, S. S. Artifact-and content-specific quality assessment for MRI with image rulers. *Medical Image Analysis* **77**, 102344 (2022).
19. Barrett, H. H., Myers, K. J., Hoeschen, C., Kupinski, M. A. & Little, M. P. Task-based measures of image quality and their relation to radiation dose and patient risk. *Physics in Medicine & Biology* **60**, R1 (2015).
20. Pryde, M., Reeve, S., Bouchie, T., Cora, E. A., Volders, D., Schmidt, M., Abdoell, M., Bowen, C., Rioux, J. & Beyea, S. The Performance of Image Quality Metrics Depends on the Diagnostic Task: A Case Study in Stroke MRI. *ISMRM* (2022).

21. Sciarra, A., Chatterjee, S., Dünwald, M., Placidi, G., Nürnberger, A., Speck, O. & Oeltze-Jafra, S. *Reference-less SSIM Regression for Detection and Quantification of Motion Artefacts in Brain MRIs* in MIDL (2022).
22. Tisdall, M. D., Hess, A. T., Reuter, M., Meintjes, E. M., Fischl, B. & van der Kouwe, A. J. Volumetric navigators for prospective motion correction and selective reacquisition in neuroanatomical MRI. *Magnetic resonance in medicine* **68**, 389–399 (2012).
23. Andronesi, O. C., Bhattacharyya, P. K., Bogner, W., Choi, I.-Y., Hess, A. T., Lee, P., Meintjes, E. M., Tisdall, M. D., Zaitzev, M. & van der Kouwe, A. Motion correction methods for MRS: experts’ consensus recommendations. *NMR in Biomedicine* **34**, e4364 (2021).
24. Slipsager, J. M., Ellegaard, A. H., Glimberg, S. L., Paulsen, R. R., Tisdall, M. D., Wighton, P., van der Kouwe, A., Marner, L., Henriksen, O. M., Law, I., *et al.* Markerless motion tracking and correction for PET, MRI, and simultaneous PET/MRI. *Plos one* **14**, e0215524 (2019).
25. Pardoe, H. R., Martin, S. P., Zhao, Y., George, A., Yuan, H., Zhou, J., Liu, W. & Devinsky, O. Estimation of in-scanner head pose changes during structural MRI using a convolutional neural network trained on eye tracker video. *Magnetic Resonance Imaging* **81**, 101–108 (2021).
26. Musa, M., Sengupta, S. & Chen, Y. MRI-Compatible Soft Robotic Sensing Pad for Head Motion Detection. *IEEE Robotics and Automation Letters* **7**, 3632–3639 (2022).
27. Laustsen, M., Andersen, M., Xue, R., Madsen, K. H. & Hanson, L. G. Tracking of rigid head motion during MRI using an EEG system. *Magnetic Resonance in Medicine* (2022).
28. Lohner, V., Lu, R., Enkirch, S., Stöcker, T., Hattingen, E. & Breteler, M. Incidental findings on 3 T neuroimaging: cross-sectional observations from the population-based Rhineland Study. *Neuroradiology* **64**, 503–512 (2022).
29. Jenkinson, M. Measuring transformation error by RMS deviation. *Studholme, C., Hill, DLG, Hawkes, DJ* (1999).
30. Peng, H., Gong, W., Beckmann, C. F., Vedaldi, A. & Smith, S. M. Accurate brain age prediction with lightweight deep neural networks. *Medical image analysis* **68**, 101871 (2021).
31. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. *Densely connected convolutional networks* in CVPR (2017), 4700–4708.

32. Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B. & Reuter, M. Fastsurfer-a fast and accurate deep learning based neuroimaging pipeline. *NeuroImage* **219**, 117012 (2020).
33. Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A. & Gorgolewski, K. J. MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PloS one* **12**, e0184661 (2017).
34. Mortamet, B., Bernstein, M. A., Jack Jr, C. R., Gunter, J. L., Ward, C., Britson, P. J., Meuli, R., Thiran, J.-P. & Krueger, G. Automatic quality assessment in structural brain magnetic resonance imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* **62**, 365–372 (2009).
35. Madan, C. R. Age differences in head motion and estimates of cortical morphology. *PeerJ* **6**, e5176 (2018).

Bibliography for the publication: FastSurfer-LIT: Lesion Inpainting Tool for Whole Brain MRI Segmentation with Tumors, Cavities and Abnormalities

1. Fischl, B. FreeSurfer. *Neuroimage* **62**, 774–781 (2012).
2. Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W. & Smith, S. M. Fsl. *Neuroimage* **62**, 782–790 (2012).
3. Ashburner, J. Computational anatomy with the SPM software. *Magnetic resonance imaging* **27**, 1163–1174 (2009).
4. Radwan, A. M., Emsell, L., Blommaert, J., Zhylka, A., Kovacs, S., Theys, T., Sollmann, N., Dupont, P. & Sunaert, S. Virtual brain grafting: Enabling whole brain parcellation in the presence of large lesions. *NeuroImage* **229**, 117731 (2021).
5. Guo, C., Ferreira, D., Fink, K., Westman, E. & Granberg, T. Repeatability and reproducibility of FreeSurfer, FSL-SIENAX and SPM brain volumetric measurements and the effect of lesion filling in multiple sclerosis. *European radiology* **29**, 1355–1364 (2019).
6. Iglesias, J. E., Billot, B., Balbastre, Y., Magdamo, C., Arnold, S. E., Das, S., Edlow, B. L., Alexander, D. C., Golland, P. & Fischl, B. SynthSR: A public AI tool to turn heterogeneous clinical brain scans into high-resolution T1-weighted images for 3D morphometry. *Science advances* **9**, eadd3607 (2023).

7. Imms, P., Clemente, A., Deutscher, E., Radwan, A. M., Akhlaghi, H., Beech, P., Wilson, P. H., Irimia, A., Poudel, G., Domínguez Duque, J. F. & Caeyenberghs, K. Exploring personalized structural connectomics for moderate to severe traumatic brain injury. *Network Neuroscience* **7**, 160–183 (Jan. 2023).
8. Zhang, S., Sun, H., Yang, X., Wan, X., Tan, Q., Li, S., Shao, H., Su, X., Yue, Q. & Gong, Q. An MRI Study Combining Virtual Brain Grafting and Surface-Based Morphometry Analysis to Investigate Contralateral Alterations in Cortical Morphology in Patients With Diffuse Low-Grade Glioma. *Journal of Magnetic Resonance Imaging* (2022).
9. Liu, D., Chen, J., Ge, H., Yan, Z., Luo, B., Hu, X., Yang, K., Liu, Y., Xiao, C., Zhang, W. & Liu, H. Structural plasticity of the contralesional hippocampus and its subfields in patients with glioma. *European Radiology* **33**, 6107–6115 (Apr. 2023).
10. Schevenels, K., Michiels, L., Lemmens, R., De Smedt, B., Zink, I. & Vandermosten, M. The role of the hippocampus in statistical learning and language recovery in persons with post stroke aphasia. *NeuroImage: Clinical* **36**, 103243 (2022).
11. Zhylka, A., Sollmann, N., Kofler, F., Radwan, A., De Luca, A., Gempt, J., Wiestler, B., Menze, B., Krieg, S. M., Zimmer, C., Kirschke, J. S., Sunaert, S., Leemans, A. & Pluim, J. P. W. Tracking the corticospinal tract in patients with high-grade glioma: clinical evaluation of multi-level fiber tracking and comparison to conventional deterministic approaches. *Frontiers in Oncology* **11**, 761169 (2021).
12. Henschel, L., Kügler, D. & Reuter, M. FastSurferVINN: Building resolution-independence into deep learning segmentation methods—A solution for HighRes brain MRI. *NeuroImage* **251**, 118933 (2022).
13. Weiss, D. A., Saluja, R., Xie, L., Gee, J. C., Sugrue, L. P., Pradhan, A., Bryan, R. N., Rauschecker, A. M. & Rudie, J. D. Automated multiclass tissue segmentation of clinical brain MRIs with lesions. *NeuroImage: Clinical* **31**, 102769 (2021).
14. Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B. & Reuter, M. Fastsurfer-a fast and accurate deep learning based neuroimaging pipeline. *NeuroImage* **219**, 117012 (2020).
15. Faber, J., Kügler, D., Bahrami, E., Heinz, L.-S., Timmann, D., Ernst, T. M., Deike-Hofmann, K., Klockgether, T., van de Warrenburg, B., van Gaalen, J., Reetz, K., Romanzetti, S., Oz, G., Joers, J. M., Diedrichsen, J., Giunti, P., Garcia-Moreno, H., Jacobi, H., Jende, J., de Vries, J., Povazan, M., Barker, P. B., Steiner, K. M., Krahe, J. & Reuter, M. CerebNet: A fast and reliable deep-learning pipeline for detailed cerebellum sub-segmentation. *NeuroImage* **264**, 119703 (2022).

16. Subramanian, S., Gholami, A. & Biros, G. Simulation of glioblastoma growth using a 3D multispecies tumor model with mass effect. *Journal of mathematical biology* **79**, 941–967 (2019).
17. Chard, D. T., Jackson, J. S., Miller, D. H. & Wheeler-Kingshott, C. A. Reducing the impact of white matter lesions on automated measures of brain gray and white matter volumes. *Journal of magnetic resonance imaging* **32**, 223–228 (2010).
18. Battaglini, M., Jenkinson, M. & De Stefano, N. Evaluating and reducing the impact of white matter lesions on brain volume measurements. *Human brain mapping* **33**, 2062–2071 (2012).
19. Magon, S., Gaetano, L., Chakravarty, M. M., Lerch, J. P., Naegelin, Y., Stippich, C., Kappos, L., Radue, E.-W. & Sprenger, T. White matter lesion filling improves the accuracy of cortical thickness measurements in multiple sclerosis patients: a longitudinal study. *BMC neuroscience* **15**, 1–10 (2014).
20. Popescu, V., Ran, N., Barkhof, F., Chard, D., Wheeler-Kingshott, C. & Vrenken, H. Accurate GM atrophy quantification in MS using lesion-filling with co-registered 2D lesion masks. *NeuroImage: Clinical* **4**, 366–373 (2014).
21. Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., Schulz, U. G., Kuker, W., Battaglini, M., Rothwell, P. M. & Jenkinson, M. BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities. *NeuroImage* **141**, 191–205 (2016).
22. Prados, F., Cardoso, M. J., Kanber, B., Ciccarelli, O., Kapoor, R., Wheeler-Kingshott, C. A. G. & Ourselin, S. A multi-time-point modality-agnostic patch-based method for lesion filling in multiple sclerosis. *Neuroimage* **139**, 376–384 (2016).
23. Schmidt, P., Pongratz, V., Küster, P., Meier, D., Wuerfel, J., Lukas, C., Bellenberg, B., Zipp, F., Groppa, S., Sämann, P. G., Weber, F., Gaser, C., Franke, T., Bussas, M., Kirschke, J., Zimmer, C., Hemmer, B. & Mührlau, M. Automated segmentation of changes in FLAIR-hyperintense white matter lesions in multiple sclerosis on serial magnetic resonance imaging. *NeuroImage: Clinical* **23**, 101849 (2019).
24. Mellerio, C., Labeyrie, M.-A., Chassoux, F., Roca, P., Alami, O., Plat, M., Naggara, O., Devaux, B., Meder, J.-F. & Oppenheim, C. 3T MRI improves the detection of transmantle sign in type 2 focal cortical dysplasia. *Epilepsia* **55**, 117–122 (2014).

25. Wattjes, M., Harzheim, M., Kuhl, C., Gieseke, J., Schmidt, S., Klotz, L., Klockgether, T., Schild, H. & Lutterbey, G. Does high-field MR imaging have an influence on the classification of patients with clinically isolated syndromes according to current diagnostic MR imaging criteria for multiple sclerosis? *American journal of neuroradiology* **27**, 1794–1798 (2006).
26. Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C. & Jenkinson, M. The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage* **80**. Mapping the Connectome, 105–124 (2013).
27. Zaretskaya, N., Fischl, B., Reuter, M., Renvall, V. & Polimeni, J. R. Advantages of cortical surface reconstruction using submillimeter 7T MEMPRAGE. *Neuroimage* **165**, 11–26 (2018).
28. Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C. & Buckner, R. L. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience* **19**, 1498–1507 (2007).
29. Tang, Z., Cabezas, M., Liu, D., Barnett, M., Cai, W. & Wang, C. LG-Net: lesion gate network for multiple sclerosis lesion inpainting in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII* **24** (2021), 660–669.
30. Clèrigues, A., Valverde, S., Salvi, J., Oliver, A. & Lladó, X. Minimizing the effect of white matter lesions on deep learning based tissue segmentation for brain volumetry. *Computerized Medical Imaging and Graphics* **103**, 102157 (2023).
31. Xie, K., Gao, L., Zhang, H., Zhang, S., Xi, Q., Zhang, F., Sun, J., Lin, T., Sui, J. & Ni, X. GAN-based metal artifacts region inpainting in brain MRI imaging with reflective registration. *Medical Physics* (2023).
32. Xiao, Y., Ashbee, W., Calhoun, V. D. & Plis, S. *Refacing Defaced MRI with PixelCNN* in *2022 International Joint Conference on Neural Networks (IJCNN)* (2022), 1–7.
33. Xing, F., Liu, X., Kuo, C.-C. J., El Fakhri, G. & Woo, J. Brain MR atlas construction using symmetric deep neural inpainting. *IEEE journal of biomedical and health informatics* **26**, 3185–3196 (2022).
34. Wang, J., Larson, K. E. & Oguz, I. Self-Supervised CSF Inpainting with Synthetic Atrophy for Improved Accuracy Validation of Cortical Surface Analyses. *arXiv preprint arXiv:2303.05777* (2023).

35. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R. & Van Gool, L. *Repaint: Inpainting using denoising diffusion probabilistic models* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), 11461–11471.
36. Dhariwal, P. & Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* **34**, 8780–8794 (2021).
37. Chung, H., Ryu, D., McCann, M. T., Klasky, M. L. & Ye, J. C. *Solving 3d inverse problems using pre-trained 2d diffusion models* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), 22542–22551.
38. Lee, S., Chung, H., Park, M., Park, J., Ryu, W.-S. & Ye, J. C. Improving 3D Imaging with Pre-Trained Perpendicular 2D Diffusion Models. *arXiv preprint arXiv:2303.08440* (2023).
39. Roy, S., K gler, D. & Reuter, M. *Are 2.5D approaches superior to 3D deep networks in whole brain segmentation?* in *International Conference on Medical Imaging with Deep Learning* (2022), 988–1004.
40. Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.-A., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., Collins, D. L., Cordier, N., Corso, J. J., Criminisi, A., Das, T., Delingette, H., Demiralp, C., Durst, C. R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K. M., Jena, R., John, N. M., Konukoglu, E., Lashkari, D., Mariz, J. A., Meier, R., Pereira, S., Precup, D., Price, S. J., Raviv, T. R., Reza, S. M. S., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H.-C., Shotton, J., Silva, C. A., Sousa, N., Subbanna, N. K., Szekely, G., Taylor, T. J., Thomas, O. M., Tustison, N. J., Unal, G., Vasseur, F., Wintermark, M., Ye, D. H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M. & Van Leemput, K. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging* **34**, 1993–2024 (Oct. 2015).
41. Baid, U., Ghodasara, S., Mohan, S., *et al.* The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. *arXiv: 2107.02314 [cs.CV]* (2021).
42. Bakas, S., Sako, C., Akbari, H., Bilello, M., Sotiras, A., Shukla, G., Rudie, J. D., Santamar a, N. F., Kazerooni, A. F., Pati, S., Rathore, S., Mamourian, E., Ha, S. M., Parker, W., Doshi, J., Baid, U., Bergman, M., Binder, Z. A., Verma, R., Lustig, R. A., Desai, A. S., Bagley, S. J., Mourelatos, Z., Morrisette, J., Watt, C. D., Brem, S., Wolf, R. L., Melhem, E. R., Nasrallah, M. P., Mohan, S., O’Rourke, D. M. & Davatzikos,

- C. The University of Pennsylvania glioblastoma (UPenn-GBM) cohort: Advanced MRI, clinical, genomics, & radiomics. *Scientific data* **9**, 453 (2022).
43. Aerts, H., Schirner, M., Dhollander, T., Jeurissen, B., Achten, E., Van Roost, D., Ritter, P. & Marinazzo, D. Modeling brain dynamics after tumor resection using The Virtual Brain. *NeuroImage* **213**, 116738 (2020).
44. Aerts, H. & Marinazzo, D. *BTC_preop* (OpenNeuro, 2018).
45. Liew, S.-L., Anglin, J. M., Banks, N. W., Sondag, M., Ito, K. L., Kim, H., Chan, J., Ito, J., Jung, C., Khoshab, N., Lefebvre, S., Nakamura, W., Saldana, D., Schmiesing, A., Tran, C., Vo, D., Ard, T., Heydari, P., Kim, B., Aziz-Zadeh, L., Cramer, S. C., Liu, J., Soekadar, S., Nordvik, J.-E., Westlye, L. T., Wang, J., Winstein, C., Yu, C., Ai, L., Koo, B., Craddock, R. C., Milham, M., Lakich, M., Pienta, A. & Stroud, A. A large, open source dataset of stroke anatomical brain images and manual lesion segmentations. *Scientific Data* **5** (Feb. 2018).
46. Baheti, B., Waldmannstetter, D., Chakrabarty, S., Akbari, M., Bilello, M., Wiestler, B., Schwarting, J., Calabrese, E., Jeffrey, R., Abidi, S., Mousa, M., Villanueva-Meyer, J., Marcus, D. S., Davathikos, C., Sotiras, A., Menze, B. H. & Bakas, S. *The Brain Tumor Sequence Registration Challenge: Establishing Correspondence between Pre-Operative and Follow-up MRI scans of diffuse glioma patients* technical report 2112.06979 (Dec. 2021).
47. Modat, M., Cash, D. M., Daga, P., Winston, G. P., Duncan, J. S. & Ourselin, S. Global image registration using a symmetric block-matching approach. *Journal of medical imaging* **1**, 024003–024003 (2014).
48. Breteler, M. M., Stöcker, T., Pracht, E., Brenner, D. & Stirnberg, R. IC-P-165: MRI in the Rhineland study: a novel protocol for population neuroimaging. *Alzheimer's & Dementia* **10**, P92–P92 (2014).
49. Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D. A., Gallagher, L., Kennedy, D. P., Keown, C. L., Keysers, C., Lainhart, J. E., Lord, C., Luna, B., Menon, V., Minshew, N. J., Monk, C. S., Mueller, S., Müller, R.-A., Nebel, M. B., Nigg, J. T., O'Hearn, K., Pelphrey, K. A., Peltier, S. J., Rudie, J. D., Sunaert, S., Thioux, M., Tyszka, J. M., Uddin, L. Q., Verhoeven, J. S., Wenderoth, N., Wiggins, J. L., Mostofsky, S. H. & Milham, M. P. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry* **19**, 659–667 (2014).

50. Di Martino, A., O'Connor, D., Chen, B., Alaerts, K., Anderson, J. S., Assaf, M., Balsters, J. H., Baxter, L., Beggiato, A., Bernaerts, S., *et al.* Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Scientific data* **4**, 1–15 (2017).
51. Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L. Whitwell, J., Ward, C., Dale, A. M., Felmlee, J. P., Gunter, J. L., Hill, D. L., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C. S., Krueger, G., Ward, H. A., Metzger, G. J., Scott, K. T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J. P., Fleisher, A. S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J. & Weiner, M. W. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging* **27**, 685–691 (Feb. 2008).
52. IXI – Information eXtraction from Images <https://brain-development.org/ixi-dataset/>. [Accessed 21-11-2025].
53. Poldrack, R., Congdon, E., Triplett, W., Gorgolewski, K., Karlsgodt, K., Mumford, J., Sabb, F., Freimer, N., London, E., Cannon, T. & Bilder, R. A phenome-wide examination of neural and cognitive function. *Scientific Data* **3** (Dec. 2016).
54. Babayan, A., Erbey, M., Kumral, D., Reinelt, J. D., Reiter, A. M. F., Röbbig, J., Schaare, H. L., Uhlig, M., Anwander, A., Bazin, P.-L., Horstmann, A., Lampe, L., Nikulin, V. V., Okon-Singer, H., Preusser, S., Pampel, A., Rohr, C. S., Sacher, J., Thöne-Otto, A., Trapp, S., Nierhaus, T., Altmann, D., Arelin, K., Blöchl, M., Bongartz, E., Breig, P., Cesnaite, E., Chen, S., Cozatl, R., Czerwonatis, S., Dambrauskaite, G., Dreyer, M., Enders, J., Engelhardt, M., Fischer, M. M., Forschack, N., Golchert, J., Golz, L., Guran, C. A., Hedrich, S., Hentschel, N., Hoffmann, D. I., Huntenburg, J. M., Jost, R., Kosatschek, A., Kunzendorf, S., Lammers, H., Lauckner, M. E., Mahjoory, K., Kanaan, A. S., Mendes, N., Menger, R., Morino, E., Nätke, K., Neubauer, J., Noyan, H., Oligschläger, S., Panczyszyn-Trzewik, P., Poehlchen, D., Putzke, N., Roski, S., Schaller, M.-C., Schieferbein, A., Schlaak, B., Schmidt, R., Gorgolewski, K. J., Schmidt, H. M., Schrimpf, A., Stasch, S., Voss, M., Wiedemann, A., Margulies, D. S., Gaebler, M. & Villringer, A. A mind-brain-body dataset of MRI, EEG, cognition, emotion, and peripheral physiology in young and old adults. *Scientific data* **6**, 1–21 (2019).
55. Malone, I. B., Cash, D., Ridgway, G. R., MacManus, D. G., Ourselin, S., Fox, N. C. & Schott, J. M. MIRIAD — Public release of a multiple time point Alzheimer's MR imaging dataset. *NeuroImage* **70**, 33–36 (2013).

56. Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C. & Buckner, R. L. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. *Journal of cognitive neuroscience* **22**, 2677–2684 (2010).
57. Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., Pop, S. C., Girard, P., Améli, R., Ferré, J.-C., Kerbrat, A., Tourdias, T., Cervenansky, F., Glatard, T., Beaumont, J., Doyle, S., Forbes, F., Knight, J., Khademi, A., Mahbod, A., Wang, C., McKinley, R., Wagner, F., Muschelli, J., Sweeney, E., Roura, E., Lladó, X., Santos, M. M., Santos, W. P., Silva-Filho, A. G., Tomas-Fernandez, X., Urien, H., Bloch, I., Valverde, S., Cabezas, M., Vera-Olmos, F. J., Malpica, N., Guttmann, C., Vukusic, S., Edan, G., Dojat, M., Styner, M., Warfield, S. K., Cotton, F. & Barillot, C. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Scientific reports* **8**, 13650 (2018).
58. Commowick, O., Kain, M., Casey, R., Ameli, R., Ferré, J.-C., Kerbrat, A., Tourdias, T., Cervenansky, F., Camarasu-Pop, S., Glatard, T., Vukusic, S., Edan, G., Barillot, C., Dojat, M. & Cotton, F. Multiple sclerosis lesions segmentation from multiple experts: The MICCAI 2016 challenge dataset. *NeuroImage* **244**, 118589 (2021).
59. Vukusic, S., Casey, R., Rollot, F., Brochet, B., Pelletier, J., Laplaud, D.-A., Sèze, J. D., Cotton, F., Moreau, T., Stankoff, B., Fontaine, B., Guillemin, F., Debouverie, M. & Clanet, M. Observatoire Français de la Sclérose en Plaques (OFSEP): A unique multimodal nationwide MS registry in France. *Multiple Sclerosis Journal* **26**. PMID: 30541380, 118–122 (2020).
60. Confavreux, C., Compston, D., Hommes, O., McDonald, W. & Thompson, A. EDMUS, a European database for multiple sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry* **55**, 671–676 (1992).
61. SHANOIR SharingNeuroImagingResources shanoir.org. [Accessed 18-10-2024].
62. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L. & Prior, F. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of digital imaging* **26**, 1045–1057 (2013).
63. Bakas, S., Sako, C., Akbari, H., Bilello, M., Sotiras, A., Shukla, G., Rudie, J. D., Flores Santamaria, N., Fathi Kazerooni, A., Pati, S., Rathore, S., Mamourian, E., Ha, S. M., Parker, W., Doshi, J., Baid, U., Bergman, M., Binder, Z. A., Verma, R., Other, N. & Davatzikos, C. Multi-parametric magnetic resonance imaging (mpMRI) scans for de novo Glioblastoma (GBM) patients from the University of Pennsylvania

- Health System (UPENN-GBM) Version 2.* <https://doi.org/10.7937/TCIA.709X-DN49> (The Cancer Imaging Archive, 2021).
64. Aerts, H., Schirner, M., Jeurissen, B., Van Roost, D., Achten, E., Ritter, P. & Marinazzo, D. Modeling brain dynamics in brain tumor patients using the virtual brain. *Eneuro* **5** (2018).
 65. Avants, B. B., Epstein, C. L., Grossman, M. & Gee, J. C. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis* **12**, 26–41 (2008).
 66. Pinaya, W. H., Tudosiu, P.-D., Dafflon, J., Da Costa, P. F., Fernandez, V., Nachev, P., Ourselin, S. & Cardoso, M. J. *Brain imaging generation with latent diffusion models in MICCAI Workshop on Deep Generative Models* (2022), 117–126.
 67. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. *High-resolution image synthesis with latent diffusion models in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), 10684–10695.
 68. Pinaya, W. H. L., Graham, M. S., Kerfoot, E., Tudosiu, P.-D., Dafflon, J., Fernandez, V., Sanchez, P., Wolleb, J., da Costa, P. F., Patel, A., Chung, H., Zhao, C., Peng, W., Liu, Z., Mei, X., Lucena, O., Ye, J. C., Tsaftaris, S. A., Dogra, P., Feng, A., Modat, M., Nachev, P., Ourselin, S. & Cardoso, M. J. Generative AI for Medical Imaging: extending the MONAI Framework. arXiv: [2307.15208](https://arxiv.org/abs/2307.15208) [eess.IV] (2023).
 69. Pollak, C., Kügler, D., Breteler, M. M. & Reuter, M. Quantifying MR head motion in the Rhineland Study—A robust method for population cohorts. *NeuroImage* **275**, 120176 (2023).
 70. Pollak, C., Kügler, D. & Reuter, M. *Estimating Head Motion from Mr-Images in 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)* (2023), 1–5.
 71. Wang, Z., Bovik, A., Sheikh, H. & Simoncelli, E. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**, 600–612 (2004).
 72. Fomin, V., Anmol, J., Desroziere, S., Kriss, J. & Tejani, A. *High-level library to help with training neural networks in PyTorch* <https://github.com/pytorch/ignite>. 2020.
 73. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D.,

- Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P. & SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17**, 261–272 (2020).
74. Rosen, A. F., Roalf, D. R., Ruparel, K., Blake, J., Seelaus, K., Villa, L. P., Ciric, R., Cook, P. A., Davatzikos, C., Elliott, M. A., *et al.* Quantitative assessment of structural image quality. *Neuroimage* **169**, 407–418 (2018).
75. Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A. & Gorgolewski, K. J. MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PloS one* **12**, e0184661 (2017).
76. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K. & Jaeger, P. F. nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation. *arXiv preprint arXiv:2404.09556* (2024).
77. Fisher, R. A. in *Breakthroughs in statistics: Methodology and distribution* 66–70 (Springer, 1970).
78. McGraw, K. O. & Wong, S. P. Forming inferences about some intraclass correlation coefficients. *Psychological methods* **1**, 30 (1996).
79. Wolleb, J., Bieder, F., Sandkühler, R. & Cattin, P. C. *Diffusion models for medical anomaly detection* in *International Conference on Medical image computing and computer-assisted intervention* (2022), 35–45.
80. Estrada, S., Kügler, D., Bahrami, E., Xu, P., Mousa, D., Breteler, M. M., Aziz, N. A. & Reuter, M. FastSurfer-HypVINN: Automated sub-segmentation of the hypothalamus and adjacent structures on high-resolucional brain MRI. *Imaging Neuroscience* **1**, 1–32 (2023).
81. Reuter, M., Tisdall, M. D., Qureshi, A., Buckner, R. L., van der Kouwe, A. J. & Fischl, B. Head motion during MRI acquisition reduces gray matter volume and thickness estimates. *Neuroimage* **107**, 107–115 (2015).

Bibliography for the publication: FastSurfer-CC: A robust, accurate, and comprehensive framework for corpus callosum morphometry

1. Unterberger, I., Bauer, R., Walser, G. & Bauer, G. Corpus callosum and epilepsies. *Seizure* **37**, 55–60 (2016).

2. Piven, J., Bailey, J., Ranson, B. J., Arndt, S., *et al.* An MRI study of the corpus callosum in autism. *American Journal of Psychiatry* **154**, 1051–1056 (1997).
3. Woodruff, P., McManus, I. & David, A. Meta-analysis of corpus callosum size in schizophrenia. *Journal of Neurology, Neurosurgery & Psychiatry* **58**, 457–461 (1995).
4. Ozturk, A., Smith, S., Gordon-Lipkin, E., Harrison, D., Shiee, N., Pham, D., Caffo, B., Calabresi, P. & Reich, D. MRI of the corpus callosum in multiple sclerosis: association with disability. *Multiple Sclerosis Journal* **16**, 166–177 (2010).
5. Jaatela, J., Nurmi, T., Vallinoja, J., Mäenpää, H., Sairanen, V. & Piitulainen, H. Altered corpus callosum structure in adolescents with cerebral palsy: connection to gait and balance. *Brain Structure and Function* **228**, 1901–1915 (2023).
6. Yang, K., Wu, Z., Long, J., Li, W., Wang, X., Hu, N., Zhao, X. & Sun, T. White matter changes in Parkinson’s disease. *npj Parkinson’s Disease* **9**, 150 (2023).
7. Grande, I., Berk, M., Birmaher, B. & Vieta, E. Bipolar disorder. *The Lancet* **387**, 1561–1572 (2016).
8. Di Paola, M., Spalletta, G. & Caltagirone, C. In vivo structural neuroanatomy of corpus callosum in Alzheimer’s disease and mild cognitive impairment using different MRI techniques: a review. *Journal of Alzheimer’s disease* **20**, 67–95 (2010).
9. Luders, E., Thompson, P. M. & Toga, A. W. The development of the corpus callosum in the healthy human brain. *Journal of Neuroscience* **30**, 10985–10990 (2010).
10. Lacalle-Aurioles, M. & Iturria-Medina, Y. Fornix degeneration in risk factors of Alzheimer’s disease, possible trigger of cognitive decline. *Cerebral Circulation-Cognition and Behavior* **4**, 100158 (2023).
11. Nowrangi, M. A. & Rosenberg, P. B. The fornix in mild cognitive impairment and Alzheimer’s disease. *Frontiers in aging neuroscience* **7**, 1 (2015).
12. Zhuang, L., Wen, W., Trollor, J. N., Kochan, N. A., Reppermund, S., Brodaty, H. & Sachdev, P. Abnormalities of the fornix in mild cognitive impairment are related to episodic memory loss. *Journal of Alzheimer’s Disease* **29**, 629–639 (2012).
13. Pinto, Y., de Haan, E. H. & Lamme, V. A. The split-brain phenomenon revisited: A single conscious agent with split perception. *Trends in cognitive sciences* **21**, 835–851 (2017).
14. Raybaud, C. The corpus callosum, the other great forebrain commissures, and the septum pellucidum: anatomy, development, and malformation. *Neuroradiology* **52**, 447–477 (2010).

15. Çavdar, S., Aydın, A. E., Algin, O. & Aydın, S. The complex structure of the anterior white commissure of the human brain: Fiber dissection and tractography study. *World Neurosurgery* **147**, e111–e117 (2021).
16. Baudo, M. C., Colombo, E. E., Pérez, M. A., Rahn, S. I., Domenech, N. C., Chuang, J., Forlizzi, V., Buosi, P., Baldoncini, M. & Bendersky, M. Three-dimensional anatomy of the anterior commissure: a Tractography and anatomical study. *World neurosurgery* **159**, e365–e374 (2022).
17. Jäncke, L., Staiger, J. F., Schlaug, G., Huang, Y. & Steinmetz, H. The relationship between corpus callosum size and forebrain volume. *Cerebral cortex (New York, NY: 1991)* **7**, 48–56 (1997).
18. Adamson, C., Beare, R., Walterfang, M. & Seal, M. Software pipeline for mid-sagittal corpus callosum thickness profile processing: automated segmentation, manual editor, thickness profile generator, group-wise statistical comparison and results display. *Neuroinformatics* **12**, 595–614 (2014).
19. Vermeulen, C. L., du Toit, P. J., Venter, G. & Human-Baron, R. A morphological study of the shape of the corpus callosum in normal, schizophrenic and bipolar patients. *Journal of anatomy* **242**, 153–163 (2022).
20. Ardekani, B., Lee, S. & Bachman, A. *Automated segmentation of corpus callosum in MRI scans: application to tracking alzheimer's disease progression in Conference Abstract: ACNS-2013 Australasian Cognitive Neuroscience Society Conference* (2013).
21. Fischl, B. FreeSurfer. *Neuroimage* **62**, 774–781 (2012).
22. Jlassi, A., Elbedoui, K., Barhoumi, W. & Maktouf, C. 3DCC-MPNN: automated 3D reconstruction of corpus callosum based on modified PNN and marching cubes. *Evolving Systems*, 1–27 (2024).
23. Chandra, A., Verma, S., Raghuvanshi, A. & Bodhey, N. K. PCcS-RAU-Net: automated parcellated corpus callosum segmentation from brain MRI images using modified residual attention U-Net. *Biocybernetics and Biomedical Engineering* **43**, 403–427 (2023).
24. Gajawelli, N., Paulli, A., Deoni, S., Paquette, N., Darakjian, D., Salazar, C., Dean, D., O'Muircheartaigh, J., Nelson, M. D., Wang, Y., *et al.* Surface-based morphometry of the corpus callosum in young children of ages 1–5. *Human Brain Mapping* **45**, e26693 (2024).
25. Platten, M., Martola, J., Fink, K., Ouellette, R., Piehl, F. & Granberg, T. MRI-based manual versus automated corpus callosum volumetric measurements in multiple sclerosis. *Journal of Neuroimaging* **30**, 198–204 (2020).

26. Brusini, I., Platten, M., Ouellette, R., Piehl, F., Wang, C. & Granberg, T. Automatic deep learning multicontrast corpus callosum segmentation in multiple sclerosis. *Journal of Neuroimaging* **32**, 459–470 (2022).
27. Huang, W., Chen, M., Lyu, G. & Tang, X. A deformation-based shape study of the corpus callosum in first episode schizophrenia. *Frontiers in Psychiatry* **12**, 621515 (2021).
28. Ardekani, B. A. & Alzheimer's Disease Neuroimaging Initiative. A new approach to symmetric registration of longitudinal structural MRI of the human brain. *Journal of Neuroscience Methods* **373**, 109563 (2022).
29. Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W. & Smith, S. M. Fsl. *Neuroimage* **62**, 782–790 (2012).
30. Raaf, N. & Westerhausen, R. Hand preference and the corpus callosum: Is there really no association? *Neuroimage: Reports* **3**, 100160 (2023).
31. Van Schependom, J., Jain, S., Cambron, M., Vanbinst, A.-M., De Mey, J., Smeets, D. & Nagels, G. Reliability of measuring regional callosal atrophy in neurodegenerative diseases. *NeuroImage: Clinical* **12**, 825–831 (2016).
32. Wong, K. K., Xu, W., Ayoub, M., Fu, Y.-L., Xu, H., Shi, R., Zhang, M., Su, F., Huang, Z. & Chen, W. Brain image segmentation of the corpus callosum by combining Bi-Directional Convolutional LSTM and U-Net using multi-slice CT and MRI. *Computer Methods and Programs in Biomedicine* **238**, 107602 (2023).
33. Cover, G., Herrera, W. G., Bento, M. P., Appenzeller, S. & Rittner, L. Computational methods for corpus callosum segmentation on MRI: a systematic literature review. *Computer methods and programs in biomedicine* **154**, 25–35 (2018).
34. Chang, J., Steyn, F., Ngo, S., Henderson, R., Guo, C., Bollmann, S., Fripp, J., Barth, M. & Shaw, T. Open-source hypothalamic-Fornix (OSHx-X) atlases and segmentation tool for 3T and 7T. *Journal of Open Source Software* **7**, 4368 (2022).
35. Greve, D. N., Billot, B., Cordero, D., Hoopes, A., Hoffmann, M., Dalca, A. V., Fischl, B., Iglesias, J. E. & Augustinack, J. C. A deep learning toolbox for automatic segmentation of subcortical limbic structures from MRI images. *Neuroimage* **244**, 118610 (2021).
36. Friedrich, P., Forkel, S. J. & de Schotten, M. T. Mapping the principal gradient onto the corpus callosum. *Neuroimage* **223**, 117317 (2020).
37. Witelson, S. F. The brain connection: the corpus callosum is larger in left-handers. *Science* **229**, 665–668 (1985).

38. Hofer, S. & Frahm, J. Topography of the human corpus callosum revisited—comprehensive fiber tractography using diffusion tensor magnetic resonance imaging. *Neuroimage* **32**, 989–994 (2006).
39. Hampel, H., Teipel, S. J., Alexander, G. E., Horwitz, B., Teichberg, D., Schapiro, M. B. & Rapoport, S. I. Corpus callosum atrophy is a possible indicator of region- and cell type-specific neuronal degeneration in Alzheimer disease: A magnetic resonance imaging analysis. *Archives of neurology* **55**, 193–198 (1998).
40. Adamson, C. L., Wood, A. G., Chen, J., Barton, S., Reutens, D. C., Pantelis, C., Velakoulis, D. & Walterfang, M. Thickness profile generation for the corpus callosum using Laplace's equation. *Human Brain Mapping* **32**, 2131–2140 (2011).
41. Lee, S. H., Yu, D., Bachman, A. H., Lim, J. & Ardekani, B. A. Application of fused lasso logistic regression to the study of corpus callosum thickness in early Alzheimer's disease. *Journal of neuroscience methods* **221**, 78–84 (2014).
42. Fraize, J., Convert, G., Leprince, Y., Sylvestre-Marconville, F., Kerdreux, E., Auzias, G., Lefèvre, J., Delorme, R., Elmaleh-Bergès, M., Hertz-Pannier, L., *et al.* Mapping corpus callosum surface reduction in fetal alcohol spectrum disorders with sulci and connectivity-based parcellation. *Frontiers in Neuroscience* **17**, 1188367 (2023).
43. Joshi, S. H., Narr, K. L., Philips, O. R., Nuechterlein, K. H., Asarnow, R. F., Toga, A. W. & Woods, R. P. Statistical shape analysis of the corpus callosum in schizophrenia. *Neuroimage* **64**, 547–559 (2013).
44. Figueira, F. F. A., Santos, V. S. d., Figueira, G. M. A. & Silva, Â. C. M. d. Corpus callosum index: a practical method for long-term follow-up in multiple sclerosis. *Arquivos de neuro-psiquiatria* **65**, 931–935 (2007).
45. Fujimori, J., Uryu, K., Fujihara, K., Wattjes, M. P., Suzuki, C. & Nakashima, I. Measurements of the corpus callosum index and fractional anisotropy of the corpus callosum and their cutoff values are useful to assess global brain volume loss in multiple sclerosis. *Multiple sclerosis and related disorders* **45**, 102388 (2020).
46. Henschel, L., Kügler, D. & Reuter, M. FastSurferVINN: Building resolution-independence into deep learning segmentation methods—A solution for HighRes brain MRI. *NeuroImage* **251**, 118933 (2022).
47. Paulsen, J. S., Langbehn, D. R., Stout, J. C., Aylward, E., Ross, C. A., Nance, M., Guttman, M., Johnson, S., MacDonald, M., Beglinger, L. J., *et al.* Detection of Huntington's disease decades before diagnosis: the Predict-HD study. *Journal of Neurology, Neurosurgery & Psychiatry* **79**, 874–880 (2008).

48. Buckner, R. <https://surfer.nmr.mgh.harvard.edu/fswiki/Buckner40Adni60Testing>. [Accessed 21-11-2025].
49. FreeSurferWiki. <https://surfer.nmr.mgh.harvard.edu/fswiki/FsAverage>. [Accessed 21-11-2025].
50. Evans, A. C., Collins, D. L., Mills, S., Brown, E. D., Kelly, R. L. & Peters, T. M. *3D statistical neuroanatomical models from 305 MRI volumes in 1993 IEEE conference record nuclear science symposium and medical imaging conference* (1993), 1813–1817.
51. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely Connected Convolutional Networks, Aug. *arXiv preprint arXiv:1608.06993* (2016).
52. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P. & SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17**, 261–272 (2020).
53. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289–300 (1995).
54. Seabold, S. & Perktold, J. *statsmodels: Econometric and statistical modeling with python* in *9th Python in Science Conference* (2010).
55. Reuter, M., Rosas, H. D. & Fischl, B. Highly accurate inverse consistent registration: a robust approach. *Neuroimage* **53**, 1181–1196 (2010).
56. Ardekani, B. A. & Bachman, A. H. Model-based automatic detection of the anterior and posterior commissures on MRI scans. *Neuroimage* **46**, 677–682 (2009).
57. Pallavaram, S., Dawant, B. M., Koyama, T., Yu, H., Neimat, J., Konrad, P. E. & D’Haese, P.-F. Validation of a fully automatic method for the routine selection of the anterior and posterior commissures in magnetic resonance images. *Stereotactic and functional neurosurgery* **87**, 148–154 (2009).
58. Liu, Y. & Dawant, B. M. *Automatic detection of the anterior and posterior commissures on MRI scans using regression forests* in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2014), 1505–1508.

59. Avants, B. B., Epstein, C. L., Grossman, M. & Gee, J. C. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis* **12**, 26–41 (2008).
60. Tustison, N. J., Cook, P. A., Holbrook, A. J., Johnson, H. J., Muschelli, J., Devenyi, G. A., Duda, J. T., Das, S. R., Cullen, N. C., Gillen, D. L., *et al.* The ANTsX ecosystem for quantitative biological and medical imaging. *Scientific reports* **11**, 9068 (2021).
61. Pollak, C., Kügler, D., Bauer, T., Rüber, T. & Reuter, M. FastSurfer-LIT: Lesion inpainting tool for whole-brain MRI segmentation with tumors, cavities, and abnormalities. *Imaging neuroscience* **3**, imag_a_00446 (2025).
62. Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D. A., Gallagher, L., Kennedy, D. P., Keown, C. L., Keyzers, C., Lainhart, J. E., Lord, C., Luna, B., Menon, V., Minshew, N. J., Monk, C. S., Mueller, S., Müller, R.-A., Nebel, M. B., Nigg, J. T., O’Hearn, K., Pelphrey, K. A., Peltier, S. J., Rudie, J. D., Sunaert, S., Thioux, M., Tyszka, J. M., Uddin, L. Q., Verhoeven, J. S., Wenderoth, N., Wiggins, J. L., Mostofsky, S. H. & Milham, M. P. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry* **19**, 659–667 (2014).
63. Di Martino, A., O’connor, D., Chen, B., Alaerts, K., Anderson, J. S., Assaf, M., Balsters, J. H., Baxter, L., Beggiato, A., Bernaerts, S., *et al.* Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Scientific data* **4**, 1–15 (2017).
64. Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L. Whitwell, J., Ward, C., Dale, A. M., Felmlee, J. P., Gunter, J. L., Hill, D. L., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C. S., Krueger, G., Ward, H. A., Metzger, G. J., Scott, K. T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J. P., Fleisher, A. S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J. & Weiner, M. W. The Alzheimer’s disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging* **27**, 685–691 (Feb. 2008).
65. Poldrack, R., Congdon, E., Triplett, W., Gorgolewski, K., Karlsgodt, K., Mumford, J., Sabb, F., Freimer, N., London, E., Cannon, T. & Bilder, R. A phenome-wide examination of neural and cognitive function. *Scientific Data* **3** (Dec. 2016).

66. Malone, I. B., Cash, D., Ridgway, G. R., MacManus, D. G., Ourselin, S., Fox, N. C. & Schott, J. M. MIRIAD — Public release of a multiple time point Alzheimer's MR imaging dataset. *NeuroImage* **70**, 33–36 (2013).
67. Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C. & Buckner, R. L. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience* **19**, 1498–1507 (2007).
68. Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C. & Buckner, R. L. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. *Journal of cognitive neuroscience* **22**, 2677–2684 (2010).
69. Koch, A., Stirnberg, R., Estrada, S., Zeng, W., Lohner, V., Shahid, M., Ehses, P., Pracht, E. D., Reuter, M., Stoecker, T., *et al.* Versatile MRI acquisition and processing protocol for population-based neuroimaging. *Nature Protocols*, 1–23 (2024).
70. Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S. W., *et al.* The Human Connectome Project: a data acquisition perspective. *Neuroimage* **62**, 2222–2231 (2012).
71. Bookheimer, S. Y., Salat, D. H., Terpstra, M., Ances, B. M., Barch, D. M., Buckner, R. L., Burgess, G. C., Curtiss, S. W., Diaz-Santos, M., Elam, J. S., *et al.* The lifespan human connectome project in aging: an overview. *Neuroimage* **185**, 335–348 (2019).
72. IXI – Information eXtraction from Images <https://brain-development.org/ixi-dataset/>. [Accessed 21-11-2025].
73. Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., Yu, T. & the scikit-image contributors. scikit-image: image processing in Python. *PeerJ* **2**, e453 (June 2014).
74. Steinbrecher, I. & Popp, A. MeshPy – A general purpose 3D beam finite element input generator <https://imcs-compsim.github.io/meshpy>. 2021.
75. Reuter, M., Wolter, F.-E. & Peinecke, N. Laplace–Beltrami spectra as ‘Shape-DNA’ of surfaces and solids. *Computer-Aided Design* **38**, 342–366 (2006).
76. Wachinger, C., Golland, P., Kremen, W., Fischl, B., Reuter, M., Initiative, A. D. N., *et al.* BrainPrint: A discriminative characterization of brain morphology. *NeuroImage* **109**, 232–248 (2015).

77. Van Schependom, J., Niemantsverdriet, E., Smeets, D. & Engelborghs, S. Callosal circularity as an early marker for Alzheimer's disease. *NeuroImage: Clinical* **19**, 516–526 (2018).
78. Yaldizli, Ö., Atefy, R., Gass, A., Sturm, D., Glassl, S., Tettenborn, B. & Putzki, N. Corpus callosum index and long-term disability in multiple sclerosis patients. *Journal of neurology* **257**, 1256–1264 (2010).

List of Publications

1. Pollak, C., Kügler, D., Breteler, M. M. & Reuter, M. Quantifying MR head motion in the Rhineland Study–A robust method for population cohorts. *NeuroImage* **275**, 120176 (2023).
2. Pollak, C., Kügler, D. & Reuter, M. *Estimating Head Motion from Mr-Images* in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)* (2023), 1–5.
3. Pollak, C., Kügler, D., Bauer, T., Rüber, T. & Reuter, M. FastSurfer-LIT: Lesion inpainting tool for whole-brain MRI segmentation with tumors, cavities, and abnormalities. *Imaging neuroscience* **3**, imag_a_00446 (2025).
4. Pollak, C., Diers, K., Estrada, S., Kügler, D. & Reuter, M. *FastSurfer-CC: A robust, accurate, and comprehensive framework for corpus callosum morphometry* 2025. arXiv: [2511.16471](https://arxiv.org/abs/2511.16471) [cs.CV].

List of Figures

1.1. Overview of the three challenges addressed in this thesis.	2
2.1. Overview of the discussed stages of a structural MRI study.	6
2.2. Simulations of strong MRI artifacts	8
2.3. The behavior of criterion functions	18
3.1. Robust alignment of motion capture	24
4.1. Interpretability analysis for our motion detection model.	30
5.1. Overview of the proposed lesion-robust whole-brain analysis pipeline .	34
5.2. Qualitative comparison of inpainting results	36
6.1. Change of corpus callosum thickness between a group and controls and Huntington's disease patients	39
6.2. Comparison of local CC thickness estimation for FastSurfer-CC and the previous state-of-the-art.	42