Aus der

Radiologischen Universitätsklinik Tübingen

Abteilung Diagnostische und Interventionelle Neuroradiologie

Impact of automated lesion segmentation in longitudinal MRI studies on the detection rate of progressive multiple sclerosis cases

Inaugural-Dissertation

zur Erlangung des Doktorgrades

der Medizin

der Medizinischen Fakultät
der Eberhard Karls Universität
zu Tübingen

vorgelegt von

Finn, Sarah

2026

Dekanin:	Professorin Dr. S. Y. Brucker
1. Berichterstatter:	Professorin Dr. U. Ernemann
2. Berichterstatter:	Professor Dr. M. C. Kowarik

Tag der Disputation: 09.03.2026

Table of contents

List of figures and tables

List of abbreviations

Abbreviation	Definition
AI	Artificial Intelligence
ANN	Artificial neural network
CI	Confidence interval
CIS	Clinically isolated syndrome
CNN	Convolutional Neural Network
CNS	Central nervous system
DAWM	Dirty appearing white matter
DL	Deep learning
FLAIR	Fluid-attenuated inversion recovery
FN	False Negative
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
IFN-β	Interferon beta
IgG	Immunoglobulin G
IT	Information technology
KI	Künstliche Intelligenz
M	Mean

ML	Machine learning
MRI	Magnetic resonance imaging
MS	Multiple sclerosis
NEDA	No evidence of disease activity
NPV	Negative Predictive Value
PDw	Proton density-weighted
PPMS	Primary progressive multiple sclerosis
PPV	Positive Predictive Value
RIS	Radiologically isolated syndrome
RRMS	Relapsing-remitting multiple sclerosis
SD	Standard deviation
SPMS	secondary progressive multiple sclerosis
T1w	T1-weighted
T2w	T2-weighted
TN	True Negative
TP	True Positive

1 Introduction

1.1 Multiple Sclerosis

1.1.1 Epidemiology

Multiple sclerosis (MS) is a chronic inflammatory autoimmune disease that affects the central nervous system (CNS) (Steinman 1996; Filippi et al. 2018). It is characterized by localized foci of lesions, particularly in the white matter, which mainly result from demyelination of nerve fibers and subsequent secondary axonal damage (Compston and Coles 2008).

The incidence and prevalence of MS have increased during the last years, especially in younger age groups, currently resulting in about 2.5 million MS cases worldwide (Walton et al. 2020). Its prevalence average rate is around 83 per 100,000, with higher rates in Europe, the United States and parts of Australia (Browne et al. 2014). A gradient can be observed here, where the number of cases increases with a greater distance from the equator. In Germany, the prevalence measured by health insurance data is between 175 and 289 cases per 100,000 insured persons, resulting in a number of 143,000 to 199,500 patients diagnosed with MS (Jakob Holstiege et al. 2017). The typical onset age of MS is patients in their 20s or 30s, with women being affected two to three times more often than men (Pugliatti et al. 2006). MS is the most common cause of permanent disability in young adults with CNS disorders (Ramagopalan and Sadovnick 2011; Runmarker and Andersen 1993). Due to the relatively young onset age and the long course of the disease, high costs for health care systems are associated with MS (Flachenecker et al. 2017).

1.1.2 Etiology and pathogenesis

Despite intensive research, the key aspects of the etiology, pathogenesis and pathophysiology of MS remain unclear.

Regarding the etiology of MS, a combination of genetic and environmental factors is considered as likely cause (Noseworthy et al. 2000; Sawcer et al. 2014). Studies have identified a genetic predisposition for MS with a tendency for familial clustering and an increased risk of developing the disease for family members of affected individuals (Hawkes and Macgregor 2009). So far, no causal genetic mutations have been found that are directly associated with the emergence of MS. However, MS patients commonly exhibit mutations in various genetic loci responsible for the immune response, including alleles of the human leukocyte antigen system, among others (Hohlfeld 2010; Dyment et al. 2004).

It is also believed that certain environmental factors including geography, vitamin D deficiency, obesity and smoking may be correlated with the development of MS disease (Ascherio 2013; Milo and Kahana 2010). Furthermore, there is a presumption that an infection with the Epstein-Barr virus in adulthood is associated with an increased risk of developing MS, but the specific mechanism behind this association is not yet understood (Ascherio 2013; Levin et al. 2005).

Although the exact cause of MS remains unknown, there is a general agreement that its pathogenesis involves the activation of autoimmune reactions targeting myelin proteins, resulting in structural and morphological changes within the CNS (Zhao et al. 2018). The disease is characterized by a T-lymphocyte-mediated autoimmune reaction, which triggers inflammatory demyelination and neurodegeneration in the CNS (Yan et al. 2019; Liu et al. 2022). It is hypothesized that autoreactive T-lymphocytes from the periphery cross the blood-brain-barrier and migrate into the CNS (Sweeney et al. 2019), where they induce an inflammatory reaction through the release of proinflammatory cytokines (Hohlfeld et al. 2016). This allows the influx of additional immune cells like macrophages and B-lymphocytes, which produce antibodies directed against proteins in the myelin sheath (Cobo-Calvo et al. 2020; Cross et al. 2020), as well as components of the humoral immune response (Sospedra and Martin 2005).

Pathologically, this inflammation manifests as focal areas of axonal demyelination, axonal conduction block and neuronal loss in the brain and spinal cord, so-called MS lesions (Filippi et al. 2018). These lesions are characterized

by limited remyelination and reactive astrocyte proliferation, leading to the formation of gliotic scars. As these permanent changes, collectively known as sclerosis, occur in multiple areas, the disease is called “Multiple Sclerosis” (Ebers 2008; Sadovnick et al. 1996). The preferred locations of the lesions are found around the ventricles, in the cortical/juxtacortical regions, in the infratentorial areas, and in the spinal cord (Fazekas et al. 1999; Thompson et al. 2018).

1.1.3 Clinical presentation and stages of progression

MS is characterized by a variable clinical presentation and a wide range of symptoms, depending on the extension and localization of the lesions.

Currently, the classification by Lublin and Reingold from 1996 (Lublin and Reingold 1996), which distinguishes three main forms of progression, is commonly used and accepted. These include relapsing-remitting MS (RRMS), which can transition into a secondary progressive form (SPMS), as well as primary progressive MS (PPMS), which is characterized by continuous progression from the onset.

At the time of initial diagnosis, approximately 85-90% of patients are diagnosed with RRMS (Burt et al. 2019; Carlström et al. 2019). In this form of the disease, symptoms occur in episodes, characterized by neurological deficits that may partially or completely resolve between the relapses. Within a span of 20 years, about 50% of patients with RRMS transition to a SPMS (Kappos et al. 2018; Kapoor et al. 2018) with a continuous progression of the symptoms without remission phases. In around 15% of patients, the disease manifests as PPMS (Rocca et al. 2017; Tsagkas et al. 2019) at the time of diagnosis, which is characterized by a progressive increase in neurodegenerative symptoms without a previous relapsing-remitting pattern. Additionally, there is a fourth category called clinically isolated syndrome (CIS) (Chung et al. 2020; Metz et al. 2017), which describes the first episode of focal neurological deficits which last at least 24 hours, caused by inflammation or demyelination. The findings in patients with a CIS do not fully meet the diagnostic criteria for a temporal and spatial dissemination of the lesions in the CNS. However, in the course of time, 40-80%

of the CIS cases convert to MS disease (Trojano et al. 2003; Fisniku et al. 2008). A typical early symptom that manifests in MS is an inflammation of the optic nerve, known as optic neuritis (Halilovic et al. 2014). In this case, patients usually experience unilateral visual impairment, often accompanied by a central scotoma and pain during eye movement (Weinshenker et al. 1989). Further early symptoms may appear as sensory disturbances, such as paresthesia or hypesthesia, as well as chronic fatigue (Compston and Coles 2008). As the disease progresses, additional symptoms usually appear, which can affect various organ systems, depending on the extent of the MS lesions. These symptoms may manifest as motor disturbances such as spastic paresis, gait disturbances and muscle weakness (Schmidt et al. 2022). Additionally, patients may suffer from cerebellar disorders such as ataxia or nystagmus (Compston and Coles 2008), autonomic dysfunctions like bladder and rectal dysfunction (Schmidt et al. 2022), or cognitive impairments and psychological symptoms (Compston and Coles 2008; Ghasemi et al. 2017).

MS is a chronic, lifelong disease that can significantly impact the quality of life for affected patients (Danelakis et al. 2018). Although the disease itself is not fatal, the occurrence of acute relapses or progressive accumulation of functional limitations often lead to an extended period of disability (Giovannoni et al. 2016). As a result, the overall life expectancy of individuals with MS is reduced by approximately 5 to 10 years compared to the general population (Basaran et al. 2022; Brønnum-Hansen et al. 2004).

1.1.4 Diagnostic principles

Due to the variety of clinical manifestations, careful diagnostic evaluation and differentiation of MS from its differential diagnoses are essential. This is based on a thorough medical history and neurological examination (Brownlee et al. 2017). In addition, further paraclinical investigations are conducted, including magnetic resonance imaging (MRI), cerebrospinal fluid analysis and neurophysiological tests.

The medical history provides initial indications for the presence of MS, with particular emphasis on reported focal neurological deficits, especially in a relapsing-remitting course, as well as the family history (Berthele and Hemmer 2021). These described symptoms can then be specifically tested and objectively assessed in the neurological examination. If a clinical suspicion of MS arises from these findings, it must be supported by paraclinical diagnostic procedures.

For this purpose, MRI is primarily used as imaging technique of choice, since it is non-invasive and the most sensitive technique to detect MS-typical morphological changes in the CNS (Afzal et al. 2022). Therefore, it plays a crucial role in both the diagnosis based on the current McDonald criteria (Thompson et al. 2018) and in monitoring disease activity (Polman et al. 2011). A variety of MRI protocols are used to highlight different morphological tissue changes seen in MS. Common multimodal brain MRI acquisitions include T1-weighted (T1w), T2-weighted (T2w), fluid-attenuated inversion recovery (FLAIR) and proton density-weighted (PDw) images (Aslani 2020; Basaran et al. 2022). Generally, MS lesions appear brighter or hyperintense compared to surrounding tissue in T2w, PDw and FLAIR images (Bakshi et al. 2001; Filippi et al. 2019). In contrast, lesions appear hypointense in T1w images, which is why old, burned-out lesions are also referred to as “black holes”, due to the axonal loss and tissue damage (Danelakis et al. 2018). Moreover, T1w images with the addition of gadolinium contrast agent are used to detect active lesions, which unlike non-active lesions, enhance contrast agent in the first four to six weeks due to acute inflammation and associated disruption of the blood-brain barrier (Traboulsee 2004; Sormani and Bruzzi 2013).

FLAIR images are particularly valuable for MS lesion detection due to their high contrast by suppressing cerebrospinal fluid signals (Gramsch et al. 2015; Rovira et al. 2015 Aug). Therefore, this sequence is most commonly used to assess the number, volume and distribution of MS lesions in both white and gray matter, a process known as segmentation (Roy et al. 2018). Morphologically, the lesions appear round to oval, relatively well-defined and with increasing disease duration, the may present as confluent spots (García-Lorenzo et al. 2013). They are predominantly located in juxtacortical areas, periventricular including the corpus

callosum, in the infratentorial region, and within the spinal cord (Fazekas et al. 1999).

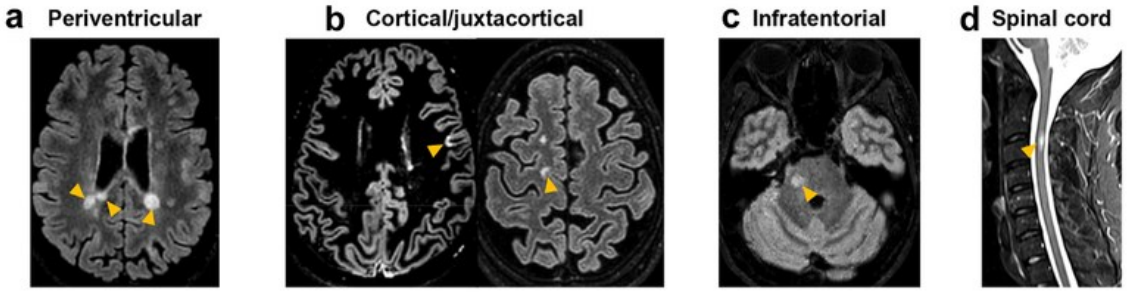

Figure 1: Typical MRI examples of MS lesions. (Reprinted from Rocca et al. 2024, under the terms of the Creative Commons Attribution License, CC BY 4.0)

The McDonald criteria, based on the assessment of spatial and temporal dissemination, were revised in 2017 (Thompson et al. 2018). Currently, spatial dissemination is classified as follows: the presence of at least one T2-hyperintense lesion in at least two of the four MS typical regions (periventricular, cortical/juxtacortical, infratentorial, spinal). In 2016, the optic nerve has been proposed as a debatable fifth MS typical location in line with the MAGNIMS consensus criteria (Filippi et al. 2016), but it has not been included in the current McDonald criteria. Temporal dissemination is defined either by the simultaneous presence of contrast-enhancing and non-enhancing lesions in the same MRI examination or the appearance of one or more new or contrast-enhancing lesions in a follow-up MRI. Applying these criteria can lead to a radiological diagnosis of MS in some MRI scans, even if the patient does not show any symptoms of MS based on their medical history or clinical examination. This condition is known as radiologically isolated syndrome (RIS) (Okuda et al. 2009). RIS may represent a preclinical stage of MS, as approximately 50% of patients with RIS are diagnosed with MS within 10 years (Lebrun-Frenay et al. 2020). Whether RIS qualifies as an MS diagnosis and if treatment should be initiated based solely on radiological findings is still an active area of research and is currently recommended only in exceptional cases (Thompson et al. 2018).

In addition to assessing lesion activity, brain volume can also be evaluated based on MRI images (Andravizou et al. 2019). Even in early stages of the disease, brain atrophy can be observed (Calabrese et al. 2007), which progresses more rapidly compared to healthy adults (Calabrese et al. 2007; Lycklama à Nijeholt, G. J. 2005). Brain volume assessment, along with lesion evaluation, serves as a measure of disease status, patient management and the evaluation of therapeutic efficacy (Zivadinov et al. 2016).

Besides MRI, further paraclinical investigations are relevant in the diagnosis of MS. Complementary tests include measuring autoantibodies in the cerebrospinal fluid and evaluating evoked potentials (Schmidt et al. 2022).

Cerebrospinal fluid analysis is primarily conducted to detect intrathecal antibody synthesis (Ruprecht and Tumani 2016), although this has a low specificity for MS, as these antibodies also occur in the context of other chronic CNS diseases. In particular, the qualitative detection of oligoclonal bands in isoelectric focusing (Dobson et al. 2013), which represents an increased synthesis of a specific IgG subfraction that can only be found in the cerebrospinal fluid and not in the serum, along with the quantitative elevation of the Immunoglobulin G (IgG) index (>0.7) according to the Reiber scheme (Wildemann et al. 2006), indicate intrathecal antibody production.

To support the diagnosis of MS, evoked potentials can be measured additionally. This test allows the detection of impaired conduction due to demyelination, which is manifested by prolonged latency time (Leocani and Comi 2008).

1.1.5 Therapy

Regarding the treatment of MS, a distinction is made between the therapy of an acute disease flare-up and a disease-modifying therapy. In addition, there are various approaches aimed at alleviating the multitude of disease symptoms (Schmidt et al. 2022).

The aim of acute relapse therapy is to achieve a rapid reduction of newly occurring neurological symptoms, typically using high-dose glucocorticoid

therapy (Grauer et al. 2001). The standard approach involves intravenous administration of 500 to 1000 mg of methylprednisolone daily for a period of 3 to 5 days, as it has an anti-inflammatory effect on both cellular and humoral immune response (Hervás-García et al. 2019). If this therapy fails or is contraindicated, an additional option to consider is plasmapheresis, which helps to reduce inflammatory factors in the blood (Weinshenker et al. 1999).

Even in MS patients without an acute relapse event, immunomodulatory therapy is frequently administered. Particularly in the past two decades, a range of new disease-modifying drugs have been approved (García-Lorenzo et al. 2013). This option is generally desirable, since preventing further relapse events has a favorable impact on the future degree of disability. However, in exceptional cases like patients with a mild course or lacking inflammatory activity, immunomodulatory therapy may be omitted. Its goal is to reduce frequency of relapses, disease activity, and delay disease progression, thereby maintaining the quality of life for affected patients (Berthele and Hemmer 2021). The choice and adjustment of therapy depend on the disease course, clinical symptoms, imaging-based presentation of disease activity and patient-specific factors. Currently, first-line options for disease-modifying therapy include interferon beta (IFN-β) preparations (Goodin et al. 2012; Rice et al. 2001), glatiramer acetate (Johnson et al. 2000; Khan et al. 2013), dimethyl fumarate (Gold et al. 2022; Kappos et al. 2008) or diroximel fumarate (Wray et al. 2022) and teriflunomide (Miller et al. 2020; O'Connor et al. 2016). These pharmaceuticals all represent immunomodulators acting at different points of intervention. If this therapy is not successful or there is high disease activity, additional highly effective medications are used for escalation therapy. In this case, second-line options include fingolimod (Cohen et al. 2010; Calabresi et al. 2014), cladribine (Giovannoni et al. 2018; Montalban et al. 2018), ozanimod (Cohen et al. 2019; Comi et al. 2019), and ponesimod (Kappos et al. 2021; Olsson et al. 2014). A further escalation step involves the use of recombinant, humanized, monoclonal antibodies such as natalizumab (O'Connor et al. 2014; Pucci et al. 2011), alemtuzumab (Riera et al. 2016; Zhang et al. 2017), ocrelizumab (Kappos et al. 2011; Hauser et al. 2020b), or ofatumumab (Hauser et al. 2020a; Bar-Or et al. 2018). Autologous stem cell

transplantation is considered as another potential treatment option (Burt et al. 2019). Due to the limited amount of research available, it is currently only performed in clinical trials for highly active forms of MS.

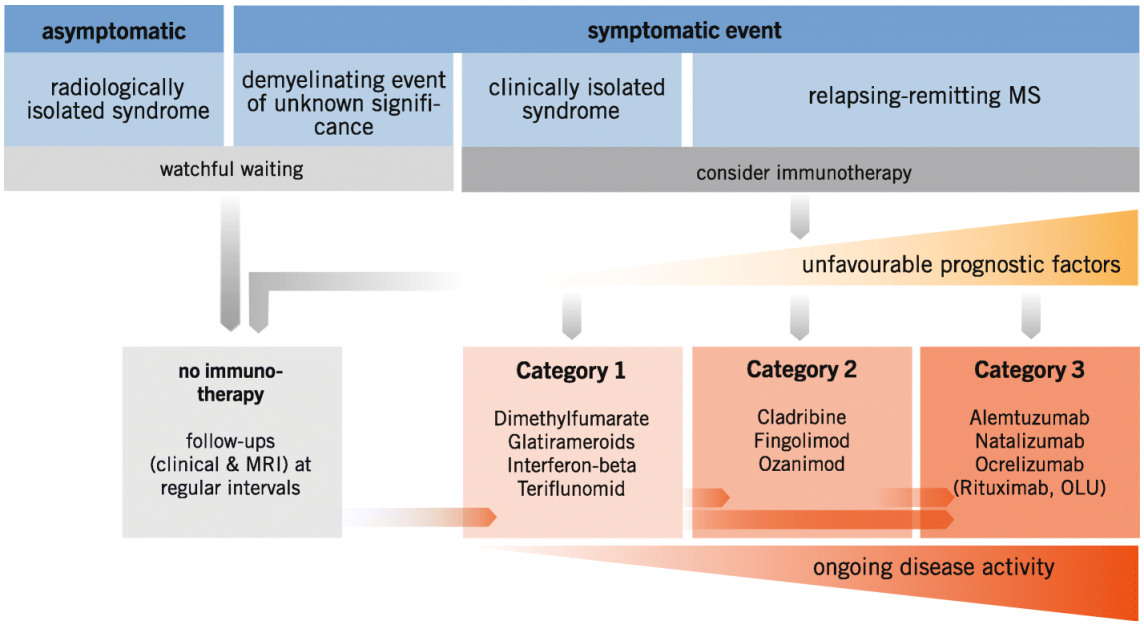

Figure 2: Treatment algorithm of the guideline. (Reprinted from Bayas et al. 2021, under the terms of the Creative Commons Attribution License, CC BY 4.0)

Since MS is currently incurable, therapy is aimed at halting disease progression (Gasperini et al. 2013). The therapeutic goal increasingly focuses on achieving “no evidence of disease activity” (NEDA) (Kappos et al. 2016; Bevan and Cree 2014), which is a classification based on clinical and radiological parameters. Clinical requirements for NEDA include being free from relapses and absence of worsening disability or cognitive decline, while radiological criteria include no increase in brain atrophy as well as no increasing lesion activity (Giovannoni et al. 2015).

Despite the use of disease-modifying therapy, the individual course of MS remains difficult to predict (Tintore et al. 2015). Therefore, intensive monitoring is essential to be able to adjust the therapy promptly in the event of disease progression (Weiner 2009).

1.2 Progress monitoring

Lesion changes, such as the presence of new or enlarged lesions, are important parameters for monitoring disease activity and assessing the effectiveness of current therapy (Calcagno et al. 2010). In particular, MRI scanning has emerged as the standard tool to monitor disease activity and the response to immunomodulatory therapy (Barkhof et al. 2011; Wattjes et al. 2015 Oct; Filippi et al. 2018). Therefore, guidelines recommend routine imaging every 6 to 12 months (Filippi et al. 2018, Hemmer B. et al. 2023). The most significant marker for monitoring chronic inflammation and disease progression in MS is the observation of lesion activity between two longitudinal MRI scans (Fahrbach et al. 2013). Increased lesion activity (McFarland et al. 1992), characterized by the appearance of new lesions measuring at least 3 mm (Rovira et al. 2015 Aug; Grahl et al. 2019) or an increase in lesion size by at least 50% compared to baseline, correlates with disease progression and is often associated with new or aggravated neurological deficits (Ouellette et al. 2018; Uher et al. 2017). In long-term studies, the occurrence of new T2-hyperintense lesions is a prognostically significant factor for predicting disease progression or the risk of increasing disability over time (Tintore et al. 2015). Hence, a precise and reproducible approach to lesion segmentation is of great interest in clinical practice. For this purpose, both manual segmentation by neuroradiologists as well as automatic methods are employed (Aslani 2020).

1.2.1 Challenges of manual MS lesion segmentation

Experienced neuroradiologists typically assess the progression of MS lesion activity by comparing chronological MRI scans in a side-by-side manner, which is considered the current gold standard (Simon et al. 2006). This process, known as segmentation, entails the manual identification and outlining of each lesion and the determination of voxel inclusion or exclusion within the lesion (Filippi et al. 1998). Nonetheless, this manual segmentation of MRI images is time-consuming, prone to errors and characterized by high intra- and inter-expert

variability (Altay et al. 2013; Egger et al. 2017), with multiple reasons accounting for this situation.

One important contributing factor is the significant heterogeneity in the quality of MRI data. Particularly when comparing images from different radiological centers, there are notable variations in scanners, field strength, protocol, sequences and resolution of the MRI acquisitions (Commowick et al. 2021). It was demonstrated that a higher magnetic field strength leads to an increased detection of lesions. (Filippi et al. 1995). Similarly, using thinner slices improves the accuracy of estimation and consequently may result in larger overall lesion volumes (Filippi et al. 1995). Additionally, the weighting of MRI images has an impact on the segmentation process. Segmented lesions on FLAIR images tend to yield larger volumes compared to T2w or PDw images (Simon et al. 2006), particularly for identifying lesions in the supratentorial region. Additionally, inadequate repositioning and the resulting difficulties in comparing follow-up images pose challenges, even when dealing with the same patient using the same scanner (Krüger et al. 2020).

Another factor to consider is the diverse characteristics of MS lesions. These can exhibit significant variations in shape, structure and location among different patients, which adds to the complexity of segmenting them, particularly in patients with a high lesion burden or when small and confluent lesions are present (Altay et al. 2013; Molyneux et al. 1999; Tan et al. 2002). An additional challenge arises from the presence of dirty appearing white matter (DAWM), which represents regions with signal intensity between that of normal-appearing white matter and MS lesions (Filippi and Rocca 2010). Distinguishing DAWM from MS lesions becomes more intricate due to their diffuse boundaries and lower contrast differences (Hindsholm et al. 2022). Moreover, depending on the specific location of the lesions, it can be challenging to differentiate lesions from other structures with similar signal strength (Mortazavi et al. 2012). This is especially evident for lesions situated near the cortex (García-Lorenzo et al. 2013), around the ventricles (Neema et al. 2009) or in the infratentorial region (Krüger et al. 2022).

Therefore, a comprehensive understanding of neuroanatomy and experience in MRI interpretation are crucial, especially due to the representation of three-dimensional brain structures in only two-dimensional coronal, sagittal or transverse slices (García-Lorenzo et al. 2013). Given that the accuracy and consistency of the results depend on the expertise of the rater (Abraham et al. 2009; Wang et al. 2017), it is essential to have experienced neuroradiologists perform MS lesion segmentation. However, outside of controlled study populations, MS MRI studies are often evaluated by radiologists with varying levels of expertise, ranging from general radiologists to neuroradiologists with specialized training (Wang et al. 2017). Numerous studies have explored the inherent variability in manual lesion segmentation performed by neuroradiologists and other trained professionals, revealing volume discrepancies ranging from 10% to 68% (Styner et al. 2008; Grimaud et al. 1996; Zijdenbos et al. 2002). Additionally, since segmentation is a subjective task (Danelakis et al. 2018), there is often a lack of agreement even among experienced neuroradiologists regarding the results (Bozsik et al. 2022).

These factors collectively contribute to the fact that MS lesion segmentation is a time-consuming and consequently expensive process (Bonacchi et al. 2022), particularly when attempting to identify newly formed lesions compared to baseline scans (Basaran et al. 2022). With advancements in technology and the increased accessibility of imaging equipment, the number of neuroradiological examinations is consistently rising (European Society of Radiology 2009 2010). Consequently, there is limited time available within the clinical routine. For example, a study conducted by Altay et al. (Altay et al. 2013) estimated a maximum of 10 minutes for a clinician to manually count lesions in an MS dataset.

In summary, automating the detection of new lesions would greatly enhance the evaluation of disease activity in patients with MS in both clinical routine and clinical trials (Ma et al. 2022). However, for widespread use, it is crucial that such a method is reliable, reproducible and efficient enough to handle the processing of large numbers of datasets (Tran et al. 2022).

1.3 Artificial intelligence in radiology

One effective way to address these challenges is by utilizing Artificial Intelligence (AI) programs. AI involves enabling computers to perform tasks that are typically associated with human intelligence (Muthukrishnan et al. 2020). The appeal of AI algorithms in the medical field lies in their ability to automate repetitive tasks, process large volumes of data at a faster rate and achieve higher levels of accuracy and reproducibility compared to humans (Hamet and Tremblay 2017).

One of the main approaches for implementing AI in radiology is Machine Learning (ML) (Afzal et al. 2022). ML is used to analyze existing data, learn from it and make decisions based on that information without explicit programming (Barile et al. 2019). In general, all ML methods can be further categorized into supervised learning, where the algorithm is trained with labeled data, and unsupervised learning, where the algorithm discovers patterns and structures within unlabeled data (Lundervold and Lundervold 2019; García-Lorenzo et al. 2013).

In unsupervised learning, the algorithm simply receives the unlabeled input data and attempts to discover categories or relationships within it without predefined criteria (Zaharchuk et al. 2018). Unsupervised ML models are applied for three main tasks: clustering, association and dimensionality reduction (Bonacchi et al. 2022). A clustering ML algorithm aims to uncover inherent groupings in the data based on similarity metrics. On the other hand, an association ML algorithm seeks to discover rules that indicate associations between variables. When the number of features in a dataset is excessively large, a dimensionality reduction ML algorithm is employed. An advantage of unsupervised methods is the fact that these algorithms do not require training, although their results may be less accurate as they do not have any prior input data to learn from (Wattjes et al. 2007). In the context of MS imaging, most unsupervised methods rely on clustering techniques to assign voxels to different classes such as tissue types like white matter, grey matter, cerebrospinal fluid, and lesion classes, based on specific features like voxel intensity (Tran et al. 2022; García-Lorenzo et al. 2013).

On the contrary, supervised methods rely on pre-labeled datasets. These datasets serve as a base for training the algorithm to establish a relationship

between input data and corresponding target values (Oren et al. 2020). The advantage of this elaborate training process is that this learned knowledge enables the algorithm to make predictions or classifications on new or unknown data. Supervised learning can be applied to classification tasks, where the output is limited to a predefined set of values, or regression tasks, where the output is a continuous value within a specific range (Bonacchi et al. 2022). Some commonly employed supervised classification methods include k-nearest neighbors (Fartaria et al. 2016), decision random forests (Jesson and Arbel; Geremia et al. 2011), support vector machines (Yamamoto et al. 2010), and more recently, artificial neural networks (ANNs) (Valverde et al. 2019).

As already mentioned, supervised machine learning includes a subset known as ANNs, which utilize interconnected nodes called artificial neurons to learn from data, perform intricate computations and make decisions based on learned patterns (Eichinger et al. 2020). The term deep learning (DL) (LeCun et al. 2015) refers to ANNs that consist of more than three individual layers, although the term is not strictly defined. While these networks were initially developed to simulate human brain activity (Lundervold and Lundervold 2019), they have limited similarities with biological neurons and essentially provide instructions on how to generate outputs from multiple inputs (Eichinger et al. 2020). The inclusion of multiple layers allows the network to grasp complex patterns and features within the data and process high-dimensional information (Zaharchuk et al. 2018). DL has recently gained significant popularity and has become the dominant form of ML due to advancements in theory, openly available computer software and the availability of hardware with sufficient computational power (Zaharchuk et al. 2018).

Convolutional Neural Networks (CNNs) are among the most popular types of ANNs. CNNs are designed with multiple layers, including one or more convolutional layers (Zaharchuk et al. 2018). Over the past decade, it has been demonstrated that CNNs are exceptionally well-suited for image analysis tasks, establishing them as a preferred choice for image-related applications (Eitel et al. 2019; Wang et al. 2018).

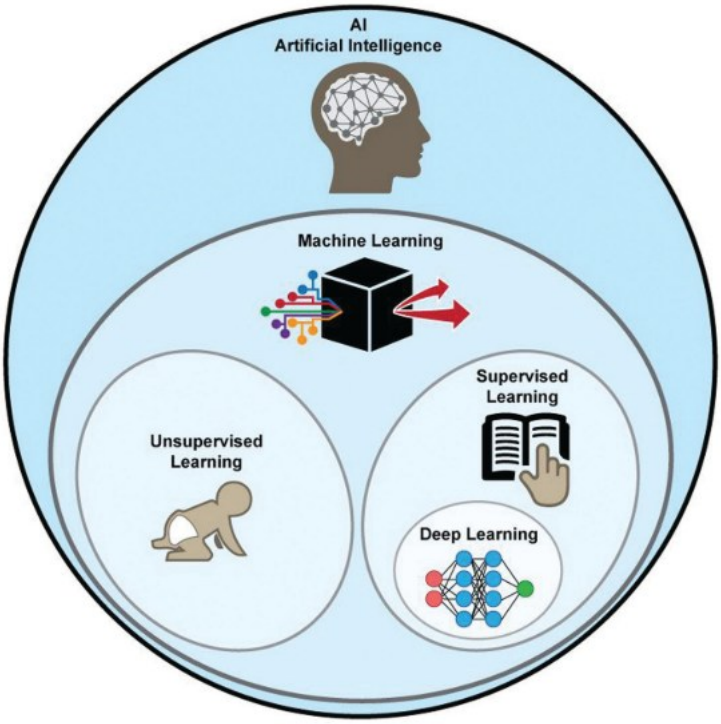

Figure 3: Artificial intelligence methods. (Reprinted from Zaharchuk et al. 2018, permission obtained through Copyright Clearance Center, License Number: 1629907-1)

1.3.1 AI-based image analysis in MS

1.3.1.1 Advantages of using AI-based image analysis programs

The application of AI-based image analysis programs for evaluating MRI images of MS patients is favored by several factors. Neuroradiology in general has a significant advantage as an early adopter of DL technologies, given to the large amount of imaging data obtained in clinical routine (Zaharchuk et al. 2018). In this context, MS is one of the medical conditions for which AI-supported image analysis is gaining importance. This is due to the fact that MS is a common disease, and the affected patients undergo regular MRI examinations (Eichinger et al. 2020), leading to a large number of MRI data. The detection of new MS lesions plays a significant role in various clinical aspects of MS: it is important for diagnosis, prognosis, as well as treatment monitoring (Brosch et al. 2022). Furthermore, the development of automated image analysis programs is also favored by the generally growing interest in AI in recent years (Eichinger et al. 2020).

The use of DL in MS diagnostics and progress monitoring focuses particularly on the identification and segmentation of MS lesions, with the aim of achieving a level of accuracy comparable to that of experienced neuroradiologists (Egger et al. 2017; Brosch et al. 2022). Advanced methods can provide rapid and reliable information on lesion burden, thereby reducing the time, effort and workload required for manual image analysis as well as reduce inter-rater variability (Valverde et al. 2017; Aslani et al. 2019; Aslani 2020; Hitziger et al. 2022).

1.3.1.2 Limitations of using AI-based image analysis programs

However, despite significant progress in the field, there is currently no established standard segmentation method for MS lesions (Danelakis et al. 2018), due to certain challenges that restrict the potential of DL in this field.

Some of these obstacles involve limited resolution of MRI images, the heterogeneous characteristics of MS lesions and the intricate structure of brain tissue, which impact numerous voxels situated at the interfaces between different tissue types (Mortazavi et al. 2012). Thus, these factors create problems in distinguishing MS lesions from the gray matter of the brain due to the overlapping patterns of intensity observed in both structures (Sahraian and Radü 2008). Moreover, certain studies demonstrated a lower precision of automated evaluation programs especially in patients with fewer lesions or smaller lesion volumes (Mortazavi et al. 2012), leading to an increased probability of false positive lesions being detected(Gabr et al. 2020; Hindsholm et al. 2022).

Additionally, numerous automated algorithms exhibit strong performance on scanners that were included in the training data, but show weak performance on data acquired from unknown scanners (Bonacchi et al. 2022). As various MRI devices are used in clinical practice, this circumstance poses a limitation for the implementation of these algorithms in clinical routine (García-Lorenzo et al. 2013). Moreover, the clinical utilization and effectiveness of these AI-based approaches are still limited by the challenges associated with obtaining consistent and reproducible measurements (Bonacchi et al. 2022). To achieve this, DL algorithms require extensive training datasets and validation across multiple datasets (Bonacchi et al. 2022). Since there is a lack of extensive

datasets accessible to the public and training AI-based segmentation methods is a time-consuming and expensive process (Aslani 2020), these methods are typically trained and tested on smaller data samples. This approach may restrict their applicability in clinical situations (Danelakis et al. 2018). Additionally, there are practical and financial concerns regarding the incorporation of AI software into current information technology (IT) systems and the availability of sufficient computing power (Bonacchi et al. 2022).

As a result, the effectiveness of AI-based methods for MS lesion segmentation continues to be a subject of ongoing research. Despite the promising potential of CNNs, their outcomes are often outperformed by experienced neuroradiologists due to the limitations mentioned above (Carass et al. 2017; Commowick et al. 2018). Having an automated technique that achieves detection and segmentation accuracy similar to that of a skilled neuroradiologist can greatly enhance the diagnostic quality by acting as a “second pair of eyes” (Hitziger et al. 2022).

This is why current investigations are focused on improving tools and techniques to enable automated analysis of MRI images within the clinical setting as a support in the diagnosis and monitoring of MS (Goldenberg 2012).

1.3.2 Use of AI-based image analysis for an untrained rater

As the quantity of medical imaging continues to rise (Eichinger et al. 2020), there is growing demand for precise and fast radiology reporting in all medical fields (Dahan et al. 2018). Moreover, the subjective nature of traditional methods used for interpreting MRI follow-up studies in MS (Danelakis et al. 2018), which depend on the rater’s skill and consistency (Abraham et al. 2009; Wang et al. 2017), can potentially be addressed through AI-based image evaluation programs.

Research has indicated that non-specialist healthcare providers with radiology training, assisted by computer-aided detection programs, have the potential to achieve comparable performance to subspecialty radiologists, for example in monitoring the progression of MS. According to a study conducted by Dahan et al. (Dahan et al. 2018), the utilization of semi-automated MS lesion segmentation

software has the potential to narrow the performance gap between experienced neuroimaging raters and inexperienced ones when monitoring MS disease. This finding is further supported by Wang et al. (Wang et al. 2017), who emphasized the advantageous impact of semi-automated assistive software in identifying new lesions for both neuroradiologists and non-neuroradiologists, with a particularly more significant effect observed in the non-neuroradiologist group.

2 Objective

Due to the technical advancements of AI-based image analysis programs, the question of their applicability in clinical practice is increasingly coming into focus, highlighting the need for further studies on this topic. Although several studies have already been published on this topic and it represents a current field of research, many questions remain unanswered. In particular, there is a gap of research regarding advantages of AI-based image analysis programs for MS-MRI evaluators with varying levels of experience in neuroradiology. To the best of my knowledge, there is currently no published study that addresses this specific issue and compares the performance between neuroradiologists and a rater without prior experience in radiology.

Therefore, this study aims to evaluate MRI images of MS patients by raters with varying levels of experience, both with and without the assistance of AI-based software. The primary objective is to evaluate how this algorithm affects the classification of the images as indicative of progressive or non-progressive course of disease. Furthermore, an additional aim of this research is to investigate the possible advantages of integrating AI-based software into the clinical routine of radiology, particularly in relation to time efficiency as well as inter- and intra-rater variability. This assessment is crucial to validate the performance of the algorithm before considering its potential application in clinical practice.

3 Materials and methods

3.1 Ethics committee

This retrospective study was planned and conducted at the University Hospital Tübingen. It was approved by the Ethics Committee at the Medical Faculty of the Eberhard Karls University and at the University Hospital Tübingen, confirmation number 662/2023BO2 and adhered to the principles of the Declaration of Helsinki.

3.2 Participants

For this study, MRI data from multiple MS patients was analyzed, which was collected at the University Hospital Tübingen as part of routine clinical practice. The inclusion criteria were the following: Patients aged 18 years or older with a confirmed diagnosis of MS according to the current McDonald criteria and a minimum of two cranial MRI scans at a minimum interval of six months, including FLAIR and T1w sequences, conducted at the University Hospital Tübingen.

The sample size calculation was based on an assumed sensitivity of 0.8 and a 95% confidence interval (CI). By setting a threshold B for the CI, the minimum sample size n can be determined using the formula $n > \left(2 \times \frac{1,96}{B}\right)^2 \times s\,(1-s)$ (Krummenauer and Kauczor 2002). For this study, a threshold B of 0.15 was chosen, resulting in a required minimum sample size of n = 110.

This research utilized a pre-pseudonymized dataset from a study involving patients with RRMS, conducted at the University Hospital Tübingen. The dataset contained the time intervals between MRI scans, with a random offset of +/- 7 days, as well as patient sex. Due to pseudonymization, information about patient age was not available. The actual dates of the study were randomly offset to ensure de-pseudonymization was not feasible. In some instances, multiple follow-up exams of the same patient were treated as separate patient cases. Ultimately, 110 patient cases were analyzed in our study, consisting of follow-up exams from 79 patients. Of these 79 patients, 55 were female and 22 were male.

3.3 MRI data

The MRI imaging was conducted on two different machines. A total of 99 patient cases were scanned using 3T MAGNETOM Vida (Siemens Healthineers, Erlangen), while 11 cases were scanned on a 1.5T MAGNETOM Avanto Fit (Siemens Healthineers, Erlangen). The imaging protocol included both FLAIR and T1w sequences. Each patient underwent two scans on the same scanner, with a minimum interval of six months between the scans.

Table 1 provides the number of scan pairs per scanner along with a summary of the sequence parameters.

Table 1: Number of scan pairs per scanner and summary of sequence parameters.

Patients (no.)	99	11
Scanner	3T MAGNETOM Vida (Siemens Healthineers, Erlangen)	1,5T MAGNETOM Avanto Fit (Siemens Healthineers, Erlangen)
FLAIR		
Echo time (ms)	0.386	0.335
Inversion time (ms)	1.8	1.8
Repetition time (ms)	5	5
Flip angle (°)	120	120
Voxel size (mm)	0.9 x 0.9 x 0.9	1.0 x 1.0 x 1.0
T1		
Echo time (ms)	0.00326	0.00296
Inversion time (ms)	1.1	0.9
Repetition time (ms)	2	2
Flip angle (°)	8	10
Voxel size (mm)	1.0 x 1.0 x 1.0	1.0 x 1.0 x 1.0

3.4 Raters and evaluation process

The MRI data was evaluated for new or enlarged MS lesions by two neuroradiologists – Rater 1, having 16 years of experience, and Rater 2 with 8

years of experience – collectively regarded as the experienced rater. The dataset was partitioned for them, with Rater 1 reviewing 53 image pairs and Rater 2 evaluating 57 pairs. Furthermore, a medical student without prior radiological experience, referred to as inexperienced rater, analyzed all data. For this rater, the assessment was split into two rounds across the two datasets (the first containing 53 cases and the second 57 cases), due to the assumption of a learning process given the lack of prior radiological experience.

All raters were blinded to the patients' clinical data and the assessment report of the other raters. The evaluation process involved comparing the chronological FLAIR images side-by-side using the program MRIcroGL. The same procedure was performed using an automated image analysis software (AIRAscore, AIRAmed GmbH), with a minimum of 90 days between those two sessions to minimize recognition bias.

The primary goal was to assess whether disease progression occurred between the two scans. Disease progression was defined solely based on the presence of MS lesions, with findings considered progressive if at least one additional lesion larger than three mm was identified or if there was a volume increase of a lesion by at least 50%.

Each rater documented the following characteristics: Firstly, they determined whether the disease course could be categorized as progressive or non-progressive. Secondly, they listed the localization and counted the number of new or enlarged lesions, if any were detected. Additionally, the duration of the assessment process for each reading was recorded and is reported as the total duration for the entire evaluation process and as average time per case. Furthermore, the untrained rater documented the sequence of the evaluations to monitor any potential learning curve.

Due to the partially high lesion burden, the number of existing unchanged lesions was not recorded.

After evaluation with and without help of an AI system, the neuroradiologist with 16 years of experience re-assessed all lesions, with knowledge of the evaluation

results of all raters and the automated results, to define the ground truth based on the raw T1 and FLAIR data. Based on this, a progressive lesion was classified as a true positive if it coincided with a ground-truth lesion.

For additional statistic evaluation, all newly developed or enlarged lesions were marked using 3D Slicer image computing software (Version 5.7.0-2024-02-29, Fedorov et al. 2012). Their volume – either the entire new lesion, or the enlarged part – was recorded in mm^3.

3.5 Software used

An automatic lesion detection system with CE certification (AIRAscore Version 2.1.0, AIRAmed GmbH) was used to assist in the reading process. This software identifies and marks lesions in a three-dimensional T2-FLAIR dataset. These lesion maps are further processed by aligning the first with the second time point and calculating the overlap of the identified lesion. Based on the extent of this overlap, lesions at the second time point are categorized as new lesions (not present in the first scan), enlarged lesions (increase >25%), or decreased lesions (decrease >25%), with different colors used to represent each category. To minimize false positives, only lesions with a volume greater than 40 µl at any point were assessed, and changes in size had to exceed 5 mm in all directions to rule out minor misalignments as the cause for change in size, using a Python script provided by AIRAmed.

3.6 Statistical Analysis

Using the ground truth as a reference, statistical parameters were gathered for each rater, both without and with the assistance of AI-based software. The selection of the statistical analyses was made in consultation with the Institute for Clinical Epidemiology and Applied Biometry at the University of Tübingen.

The analysis was split into two parts: one focusing on statistical analysis at the case level, examining whether a case was identified as progressive or not, and another segment focusing on the lesions themselves, assessing the number of

identified lesions. Additionally, a statistical analysis of the evaluation duration and the volumes of progressive lesions was conducted.

All statistical analyses were performed using IBM SPSS Statistics (Version 29.0.2.0; IBM Corp., Armonk, NY, USA). Descriptive statistics were used to summarize the data. A p-value of 0.05 was used to evaluate the significance in the statistical tests.

3.6.1 *Statistical analysis on case level*

The first part of the statistical analysis was performed on a case level. The following statistical parameters were determined for the experienced and inexperienced rater, each with and without software assistance:

- The total number of cases either categorized as "progressive" or "not progressive".
- True Positive (TP): Cases correctly identified as progressive.
- True Negative (TN): Cases correctly identified as not progressive.
- False Positive (FP): Cases incorrectly identified as progressive.
- False Negative (FN): Cases incorrectly identified as not progressive.
- Sensitivity: measures the proportion of actual progressive cases that were correctly identified.

  Formula: $Sensitivity = \frac{TP}{TP+FN}$
- Specificity: measures the proportion of actual non-progressive cases that were correctly identified, also referred to as precision.

  Formula: $Specificity = \frac{TN}{TN+FP}$
- False Positive Rate (FPR): The rate at which non-progressive cases were incorrectly identified as progressive.

  Formula: $FPR = \frac{FP}{FP+TN}$
- False Negative Rate (FNR): The rate at which progressive cases were incorrectly identified as not progressive.

  Formula: $FNR = \frac{FN}{FN+TP}$

- Positive Predictive Value (PPV): The proportion of cases identified as progressive, that were truly progressive.

  Formula: $PPV = \frac{TP}{TP+FP}$
- Negative Predictive Value (NPV): The proportion of cases identified as not progressive, that were truly not progressive.

  Formula: $NPV = \frac{TN}{TN+FN}$
- Accuracy: The proportion of total cases, that were correctly classified.

  Formula: $Accuracy = \frac{TP+TN}{Total\ cases}$
- F1 Score: This metric evaluates predictive performance by combining precision and recall. An F1 score of 1.0 signifies perfect precision and recall, while a score of 0 is the lowest value possible.

  Formula: $F1 = 2x\frac{precision\ x\ recall}{precision+recall} = \frac{2\ TP}{2\ TP+\ FP+FN}$
- Cohen's kappa coefficient (Cohen's κ) (Cohen 1960): This is as a measure of interrater agreement between the experienced and inexperienced raters, each using the software and not using the software. Cohen's Kappa ranges from -1.0 to 1.0. Interpretation (Landis and Koch 1977): κ < 0 = poor agreement, 0 – 0.20 = slight agreement, 0.21–0.40 = fair agreement, 0.41–0.60 = moderate agreement, 0.61–0.80 = substantial agreement, 0.81–1.00 = (almost) perfect agreement.

  Formula: κ = $\frac{p_0 - p_c}{1 - p_c}$ (p_0 = measured agreement value, p_c = randomly expected agreement)

The number of cases in each category (progressive or not progressive) was tabulated. Sensitivity, specificity, FPR, FNR, PPV, NPV and accuracy were calculated with 95% CIs. The 95% CIs were used to estimate whether differences in statistical parameters between evaluations without and with automated software were statistically significant.

These metrics provide a comprehensive evaluation of the diagnostic method's performance, highlighting its ability to correctly identify both progressive and non-progressive cases. The sensitivity and specificity indicate the method's effectiveness in detecting true cases, while the predictive values offer insight into

the reliability of positive and negative test results. The F1 score balances the method's precision and recall, providing an overall measure of its diagnostic accuracy.

3.6.2 *Statistical analysis on lesion level*

The second part of the statistical analysis was performed at the lesion level, evaluating both the experienced and inexperienced rater, each with and without software assistance. For monitoring disease activity in clinical practice, the count of new or enlarged lesions is an important parameter. Therefore, the following statistical parameters were assessed on a lesion level:

- The total number of lesions categorized as "progressive".
- TP: Lesions correctly identified as progressive.
- FP: Lesions incorrectly identified as progressive.
- FN: Lesions incorrectly identified as not progressive.
- Sensitivity
- PPV
- F1 score
- Cohen's κ

The number of lesions in each category (new or enlarged) was tabulated. Due to the high number of lesions in some cases, non-progressive lesions were not counted. Therefore, 95% CIs to assess statistical significance could not be calculated at the lesion level.

3.6.3 *Statistical analysis of lesion volume*

For the data classified as progressive, new or enlarged lesions were identified using 3D Slicer image computing software to measure their volume, recorded in mm^3. It was noted whether the lesion was new or had increased in size. For enlarged lesions, the increase in size was documented. Additionally, it was recorded whether both experienced and inexperienced raters identified the lesion

as progressive, both with and without using the AI-based software, and the results were compared with the lesions marked as progressive in the software overlay itself.

The statistical analysis was conducted separately for new and enlarged lesions. For new lesions, the analysis assessed whether there were size differences in lesions identified without and with AI-based software, based on mean values and their 95% CIs. The data for enlarged lesions was not normally distributed, so it was described using the median and interquartile range. Additionally, a Wilcoxon test was performed to evaluate if there was a statistically significant difference in volumes of lesions identified as progressive compared to those not identified.

3.6.4 *Statistical analysis of evaluation duration*

The experienced raters only documented the total time for evaluating the cases, not the time for individual cases. Therefore, a statistical comparison was only conducted for the inexperienced rater. Since the inexperienced rater had no prior radiological experience, an initial learning effect could not be ruled out. Therefore, an initial assessment of the duration was conducted using a paired bar chart, assuming that outliers with particularly long evaluation times would be found in the beginning. The subsequent statistical analyses were then performed both with and without the initial outliers. Additionally, the evaluation process for the inexperienced rater was divided into two rounds (the first with 53 cases and the second with 57 cases). The assessment of whether there was a significant difference in evaluation duration between round one and round two was based on mean values and their 95% CIs. Additionally, a Pearson correlation coefficient was used to evaluate the correlation between the assessment duration of round one and two. To compare evaluation duration with and without software support across all 110 cases, a paired t-test was conducted, and Cohen’s d was calculated to determine effect size.

4 Results

This study aimed to evaluate MRI images of MS patients to determine whether disease progression has occurred, based on the presence of new or enlarged lesions. The assessment was conducted by raters with varying levels of radiological experience, both with and without the support of AI-based software.

The objective was to examine how AI affects the classification of these images as either progressive or non-progressive, if it influences the time required for the evaluation process, and whether it has an impact on inter-rater and intra-rater variability.

For this purpose, a retrospective study was conducted using MRI data of 79 MS patients split into 110 cases at the University Hospital Tübingen. Evaluation of new or enlarged lesions was performed by both an experienced and inexperienced rater, each time with and without the assistance of an AI-based software. The collected data included assessments of progression at both case and lesion level, as well as the total evaluation duration and time per case. Additionally, to evaluate whether the performance of the software was dependent on the size of progressive lesions, the volume of these lesions was determined.

4.1 Results on case level

Monitoring disease activity and the effectiveness of the current therapy primarily relies on the detection of new or enlarged lesions in MRI images (Calcagno et al. 2010). Therefore, the first part of the statistical analysis was conducted at the case level, focusing on whether the case was classified as disease-progressive or non-progressive based on the occurrence of new or enlarged lesions. Out of a total of 110 cases of 79 patients, 44 cases were determined to be disease-progressive and 66 as non-progressive according to the ground truth established by a neuroradiologist with 16 years of experience.

4.1.1 *Results of the experienced rater*

The experienced rater, without aid of the automated program, classified 35 cases as progressive. Of these, 31 were TP (correctly identified as progressive) compared to the ground truth and 4 cases were FP (incorrectly identified as progressive). Additionally, there were 13 FN, meaning these cases had new or enlarged lesions according to the ground truth that were not detected, and 62 cases were TN (correctly identified as non-progressive). This resulted in a sensitivity of 0.70 (95% CI: 0.55 – 0.82) with a FNR of 0.3, and a specificity of 0.94 (95% CI: 0.85 – 0.98) with a FPR of 0.06. Additional statistical measures included a PPV of 0.89 (95% CI: 0.75 – 0.95), a NPV of 0.83 (95% CI: 0.75 – 0.88), an accuracy of 0.85 (95% CI: 0.76 – 0.91) and an F1 score of 0.79.

With automated assistance, the experienced rater classified 31 cases as progressive. Of these, 29 were TP, 2 FP, 15 FN and 64 TN. Based on these figures, the sensitivity was calculated to be 0.66 (95% CI: 0.50 – 0.80) with a FNR of 0.34, while the specificity was 0.97 (95% CI: 0.90 – 0.99) with a FPR of 0.03. Further metrics were a PPV of 0.94 (95% CI: 0.79 – 0.98), the NPV of 0.81 (95% CI: 0.74 – 0.87), an accuracy of 0.79 (95% CI: 0.70 – 0.86) and an F1 score of 0.72.

The significance of changes in statistical parameters with and without the assistance of AI-based software was assessed using the 95% CIs. Overall, the analysis showed that there were no statistically significant differences in sensitivity, specificity, PPV, NPV or accuracy when using the AI-based program compared to not using it, as suggested by the overlapping 95% CIs for these parameters.

4.1.2 *Results of the inexperienced rater*

Without assistance from the automated program, the inexperienced rater classified 48 cases as disease progressive. Of these, 42 were TP according to the ground truth, and 6 were FP. Consequently, there were 2 FN and 60 TN. From this data, the sensitivity was calculated to be 0.95 (95% CI: 0.85 – 0.99) with a FNR of 0.05, and the specificity was 0.91 (95% CI: 0.81 – 0.97) with a FPR

of 0.09. Further results included a PPV of 0.8 (95% CI: 0.77 – 0.94), a NPV of 0.97 (95% CI: 0.89 – 0.99), an accuracy of 0.93 (95% CI: 0.86 – 0.97), and an F1 score of 0.91.

With automated assistance, the inexperienced rater classified 43 cases as progressive. Of these, 41 were TP and 2 were FP. Given that there were 44 progressive cases according to the ground truth, this resulted in 3 FN and 64 TN. From this, the sensitivity was calculated to be 0.93 (95% CI: 0.81 – 0.99) with a FNR of 0.07, and the specificity was 0.97 (95% CI: 0.90 – 0.99) with a FPR of 0.03. Additionally, the PPV was 0.95 (95% CI: 0.84 – 0.99), the NPV was 0.96 (95% CI: 0.88 – 0.99), the accuracy was 0.95 (95% CI: 0.90 – 0.99), and the F1 score was 0.94.

We used 95% CIs to evaluate the significance of variations in statistical parameters, both with and without the automated program's assistance. The overlapping of the 95% CIs for sensitivity, specificity, PPV, NPV and accuracy indicated that the differences in these parameters were not statistically significant.

4.1.3 *Comparison of Statistical Analysis: Experienced vs. Inexperienced rater*

The statistical comparison between the experienced and inexperienced rater, both with and without automated assistance, was evaluated using 95% CIs.

Without the support of the automated assessment program, the sensitivity of the inexperienced rater (0.95, 95% CI: 0.85 – 0.99) was notably higher than that of the experienced rater (0.70, 95% CI: 0.55 – 0.82), with no overlap in 95% CIs, indicating a statistically significant advantage for the inexperienced rater. In contrast, the experienced rater had a slightly higher specificity (0.94, 95% CI: 0.85 – 0.98) compared to the inexperienced rater (0.91, 95% CI: 0.81 – 0.97), but the overlapping 95% CIs suggested that this difference was not statistically significant. The PPVs were close (Experienced: 0.89 vs. Inexperienced: 0.88), with overlapping 95% CIs (Experienced: 0.75 – 0.95 vs. Inexperienced: 0.77 –

0.94), indicating no statistically significant difference between the two. The inexperienced rater had a higher NPV (0.97, 95% CI: 0.89 – 0.99) compared to the experienced rater (0.83, 95% CI: 0.75 – 0.88), with non-overlapping 95% CIs, suggesting a statistically significant advantage for the inexperienced rater. Lastly, the accuracy of the inexperienced rater (0.93, 95% CI: 0.86 – 0.97) exceeded that of the experienced rater (0.85, 95% CI: 0.76 – 0.91). The 95% CIs overlapped slightly, meaning the difference was not statistically significant.

With automated assistance, the sensitivity of the inexperienced rater (0.93, 95% CI:0.81 – 0.99) was significantly higher than that of the experienced rater (0.66, 95% CI: 0.50 – 0.80), suggested by the fact that the 95% CIs did not overlap. Both raters had identical specificity values (0.97) with overlapping 95% CIs (0.90 – 0.99), showing no statistically significant difference. The PPVs were very close (Experienced: 0.94 vs. Inexperienced: 0.95), and their overlapping 95% CIs (Experienced: 0.79 – 0.98, Inexperienced: 0.84 – 0.99) suggested no statistically significant difference. The NPV was significantly higher for the inexperienced rater (0.96, 95% CI: 0.74 – 0.87) compared to the experienced rater (0.81, 95% CI: 0.88 – 0.99), since the 95% CIs did not overlap. Additionally, the accuracy was higher for the inexperienced rater (0.95) than for the experienced rater (0.79), with non-overlapping 95% CIs (Experienced: 0.70 – 0.86, Inexperienced: 0.90 – 0.99), indicating a statistically significant difference in favor of the inexperienced rater.

4.1.4 *Comparison of interrater variability using Cohen's Kappa*

At the case level, Cohen’s κ was calculated to determine the degree of agreement between the experienced and inexperienced rater, both with and without assistance of the AI-based assessment software. Without automated support, the unweighted Cohen’s κ was κ = 0.56, indicating moderate agreement. With help of the automated program, Cohen’s κ increased to κ = 0.68, reflecting improved agreement.

Since the cases were divided between the two experienced raters, with Rater 1 reviewing 53 image pairs and Rater 2 evaluating 57 pairs, Cohen’s κ was

additionally calculated separately for both rounds. In the first round which compared the interrater variability between Rater 1 and the inexperienced rater, the unweighted Cohen's κ was κ = 0.52 without automated support, and κ = 0.70 with evaluation software. In the second round, when comparing Rater 2 with the inexperienced rater, the unweighted Cohen's κ was κ = 0.58 without support and increased to κ = 0.66 with the use of AI-based software. Overall, the use of AI-based software led to an improvement in agreement from moderate to substantial in both rounds, with the effect being more pronounced in the first round (with experienced Rater 1) than in the second round (with experienced Rater 2).

4.1.5 *Results of the AI-based program*

The evaluation program itself classified 37 cases as disease-progressive, of which 29 were TP and 8 were FP. This resulted in 58 TN and 15 FN. From these figures, a sensitivity of 0.66 (95% CI: 0.50 – 0.78) with a FNR of 0.34 and a specificity of 0.88 (95% CI: 0.78 – 0.95) with a FPR of 0.12 were calculated. Also, the PPV was 0.78 (95% CI: 0.65 – 0.88), the NPV was 0.79 (95% CI: 0.72 – 0.86), the accuracy was 0.79 (95% CI: 0.70 – 0.86), and the F1 score was 0.72.

4.1.6 *Key message*

In conclusion, at the case level, no significant differences in statistical parameters were observed for either rater when using the automated program compared to the assessment without AI-based support. Both with and without the software, the inexperienced rater demonstrated higher sensitivity and NPV than the experienced rater. Other statistical parameters showed no statistically significant differences between the experienced and inexperienced rater when comparing the evaluation rounds with and without automated assistance. Cohen's κ indicated moderate agreement (κ = 0.56) without automated support, which improved to substantial agreement (κ = 0.68) with software support. This enhanced agreement was observed in both rounds with the software, with a more pronounced effect in the first round (Rater 1).

Table 2: Statistical parameters of the analysis on the case level.

	Experienced Rater without automated assistance	Experienced Rater with automated assistance	Inex-perienced Rater without automated assistance	Inex-perienced Rater with automated assistance	Automated program
cases categorized as progressive	35	31	48	43	37
True positive	31	29	42	41	29
True negative	62	64	60	64	58
False positive	4	2	6	2	8
False negative	13	15	2	3	15
Sensitivity	0.70 (95% CI: 0.55 - 0.82)	0.66 (95% CI: 0.50 - 0.80)	0.95 (95% CI: 0.85 - 0.99)	0.93 (95% CI: 0.81 - 0.99)	0.66 (95% CI: 0.50 - 0.78)
Specificity	0.94 (95% CI: 0.85 - 0.98)	0.97 (95% CI: 0.90 - 0.99)	0.91 (95% CI: 0.81 - 0.97)	0.97 (95% CI: 0.90 - 0.99)	0.88 (95% CI: 0.78 - 0.95)
False positive rate	0.06	0.03	0.09	0.03	0.12
False negative rate	0.30	0.34	0.05	0.07	0.34
Positive prediction value	0.89 (95% CI: 0.75 - 0.95)	0.94 (95% CI: 0.79 - 0.98)	0.88 (95% CI: 0.77 - 0.94)	0.95 (95% CI: 0.84 - 0.99)	0.78 (95% CI: 0.65 - 0.88)
Negative prediction value	0.83 (95% CI: 0.75 - 0.88)	0.81 (95% CI: 0.74 - 0.87)	0.97 (95% CI: 0.89 - 0.99)	0.96 (95% CI: 0.88 - 0.99)	0.79 (95% CI: 0.72 - 0.86)
Accuracy	0.85 (95% CI: 0.76 - 0.91)	0.85 (95% CI: 0.76 - 0.91)	0.93 (95% CI: 0.86 - 0.97)	0.95 (95% CI: 0.90 - 0.99)	0.79 (95% CI: 0.70 - 0.86)
F1 Score	0.79	0.77	0.91	0.94	0.72

4.2 Results on lesion level

The second part of the statistical analysis was conducted at the lesion level. For this, the number of new lesions or those that had increased in size by at least 50% in disease-progressing cases was counted. Among the 44 cases with disease progression, there were a total of 122 progressive lesions, according to the ground truth results. Since the number of existing stable lesions was not recorded, no 95% CIs could be calculated at the lesion level to assess the significance of the statistical parameters.

4.2.1 *Results of the experienced rater*

The experienced rater, without the aid of the AI-based program, identified a total of 75 lesions as new or enlarged. Of these, 59 lesions were TP, meaning they were indeed new or had increased in size, and 16 were FP. 63 lesions were FN, meaning they were new or had increased in size but were not identified. These values resulted in a sensitivity of 0.48, a PPV of 0.79 and an F1 score of 0.60.

With the help of the evaluation software, the experienced rater classified a total of 67 lesions as progressive. Among these, 65 were TP, 2 were FP, and 57 were FN. Based on these results, a sensitivity of 0.53, a PPV of 0.97 and an F1 score of 0.69 were derived.

4.2.2 *Results of the inexperienced rater*

The inexperienced rater identified a total of 114 lesions as new or enlarged without the aid of the AI-based program. According to the ground truth, 95 of these lesions were TP, while 19 were FP and 27 were FN. This resulted in a sensitivity of 0.78, a PPV of 0.83, and an F1 score of 0.81.

Using the automated software, the inexperienced rater identified a total of 117 lesions as progressive. Of these, 104 were TP, 13 were FP, and 18 were FN based on the ground truth. From these figures, a sensitivity of 0.85, a PPV of 0.89 and an F1 score of 0.87 were calculated.

4.2.3 *Comparison of statistical parameters: Experienced vs. Inexperienced rater*

Overall, the performance for both raters improved while using the AI-based software. Both with and without the software assistance, the inexperienced rater demonstrated higher sensitivity than the experienced rater. With the help of the software, sensitivity increased for both raters, with a greater increase observed for the inexperienced rater (from 0.78 to 0.85) compared to the experienced rater (from 0.48 to 0.53). For the PPV, the experienced rater showed a higher PPV with software assistance (0.97) compared to without (0.79), while the inexperienced rater's PPV was high in both cases (0.83 without and 0.89 with software). The F1 score increased for both raters when using the software, with the inexperienced rater achieving a higher score both with (0.87) and without (0.81) the software compared to the experienced rater (with software: 0.69, without: 0.60).

4.2.4 *Comparison of interrater variability using Cohen's Kappa*

Cohen's κ was calculated to assess the agreement between the experienced and inexperienced rater at the lesion level. Without aid of the automated program, the unweighted Cohen's κ was $\kappa = 0.61$, and the linear weighted Cohen's κ was also $\kappa = 0.61$, indicating substantial agreement. With automated support, the unweighted Cohen's κ increased to $\kappa = 0.64$ and the linear weighted Cohen's κ to $\kappa = 0.62$, reflecting a slight improvement in agreement.

Additionally, Cohen's κ was calculated separately for each round to assess interrater agreement between the inexperienced rater and each experienced rater separately. In the first round, with experienced rater 1 (who reviewed 53 image pairs), the unweighted Cohen's κ was $\kappa = 0.58$ and the linear weighted Cohen's κ was $\kappa = 0.54$ without the automated evaluation tool, both indicating moderate agreement. With software assistance, these values were $\kappa = 0.57$ (unweighted) and $\kappa = 0.55$ (weighted), still indicating moderate agreement. In the second round, with experienced rater 2 (who reviewed 57 image pairs), unweighted Cohen's κ was $\kappa = 0.63$ and linear weighted Cohen's κ was $\kappa = 0.65$

without software, indicating substantial agreement. When using the assessment tool, these values increased to κ = 0.70 (unweighted) and κ = 0.68 (weighted), still reflecting a substantial but increased agreement.

4.2.5 *Results of the AI-based program*

The program identified 72 lesions as new or enlarged. Out of these, 43 were TP, 29 were FP, and 79 were FN. From these findings, the sensitivity was determined to be 0.35, the PPV was 0.6 and the F1 score was 0.44.

4.2.6 *Key message*

At the lesion level, statistical parameters indicated that sensitivity, PPV and F1 scores improved for both raters when using the automated software. The inexperienced rater achieved higher sensitivity and F1 scores than the experienced rater, both with and without software support. The experienced rater, however, showed a notable increase in PPV from 0.79 without to 0.97 with software support, surpassing the inexperienced rater's PPV with automated support (0.89). Cohen's κ analysis showed a slight overall improvement in interrater agreement with software support, particularly evident between the inexperienced rater and experienced rater 2, who reached a substantial agreement level, while agreement between the inexperienced rater and experienced rater 1 remained moderate.

Table 3: Statistical parameters of the analysis on the lesion level.

	Experienced Rater without automated assistance	Experienced Rater with automated assistance	Inex-perienced Rater without automated assistance	Inex-perienced Rater with automated assistance	Automated program
Lesions categorized as progressive	75	67	114	117	72
True positive	59	65	95	104	43
False positive	16	2	19	13	29
False negative	63	57	27	18	79
Sensitivity	0.48	0.53	0.78	0.85	0.35
Positive prediction value	0.79	0.97	0.83	0.89	0.60
F1 Score	0.60	0.69	0.81	0.87	0.44

4.3 Lesion Volume

The statistical analysis of progressive lesions was conducted separately for new and enlarged lesions. Out of a total of 122 progressive lesions, there were 99 new and 23 enlarged lesions according to the ground truth.

4.3.1 *Volume of new lesions*

Out of the 122 progressive lesions, 99 were new lesions. The following section describes the results of the lesions identified as progressive and non-progressive for each rater both without and with AI-based software assistance.

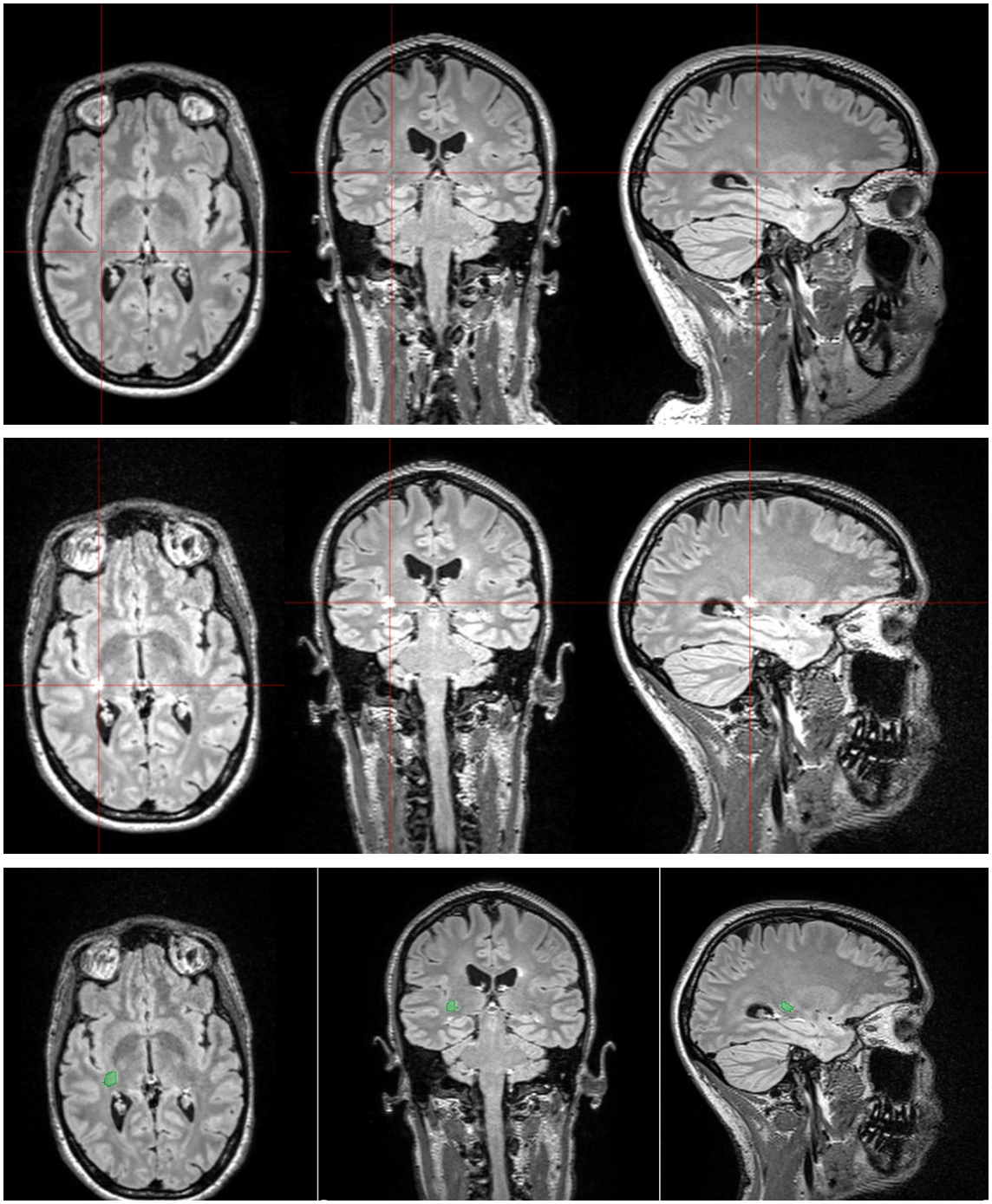

Figure 4: Example of a new MS lesion. The first image shows timepoint 1, the second shows timepoint 2, and the third displays timepoint 2 with color-coded segmentation of the new lesion using 3D Slicer.

4.3.1.1 Results of the experienced rater

Without using the automated software, the experienced rater identified 49 out of the 99 new lesions as progressive, which results in a sensitivity of 0.49. The remaining 50 new lesions were not classified as progressive and therefore were considered as FN. The statistical analysis showed that the TP lesions had an average volume of 93.06 mm^3 (95% CI: 49.32 – 136.81 mm^3) with a standard deviation (SD) of 152.30 mm^3, whereas the FN lesions had an average volume of 32.15 mm^3 (95% CI: 16.38 – 47.92 mm^3) with an SD of 55.49 mm^3. The difference in average volume between detected and undetected lesions was 60.91 mm^3. An independent t-test ($t(97) = 2.654$, $p = 0.009$) showed that the undetected lesions were significantly smaller than those that were correctly identified, which was also suggested by the non-overlapping 95% CIs.

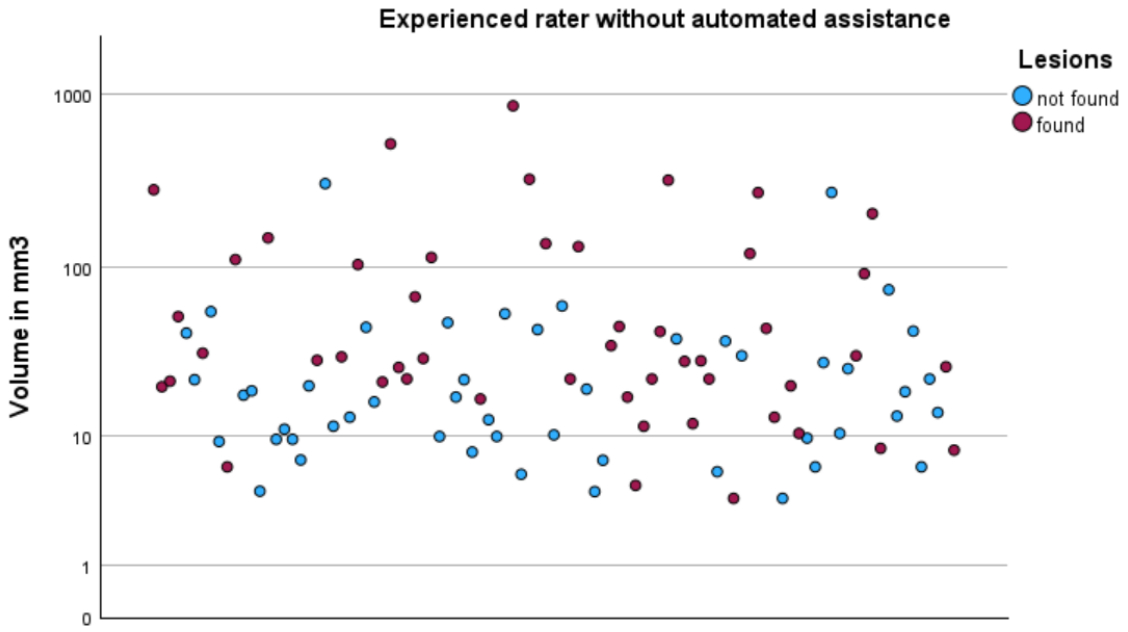

Figure 5: Detection of new lesions by the experienced rater without automated assistance. The y-axis shows lesion volume in mm^3 on a logarithmic scale. Lesions identified by the rater are marked in red ("found"), while missed lesions are shown in blue ("not found").

With assistance of the automated program, the experienced rater classified 53 of the 99 new lesions as progressive. This corresponds to a sensitivity of 0.54. Consequently, the remaining 46 new lesions were classified as FN. The statistical analysis revealed the following results: the lesions identified as progressive had

an average volume of 91.57 mm^3 (95% CI: 50.49 – 132.65 mm^3) with an SD of 149.04 mm^3. In contrast, the lesions not identified as progressive had an average volume of 28.58 mm^3 (95% CI: 14.35 – 42.80 mm^3) with an SD of 47.90 mm^3. The volume difference between the detected and undetected lesions was 62.99 mm^3. An independent t-test ($t(97) = 2.745$, $p = 0.007$) indicated that the lesions classified as progressive were significantly larger in volume than those classified as non-progressive, a finding also supported by the non-overlapping 95% CIs.

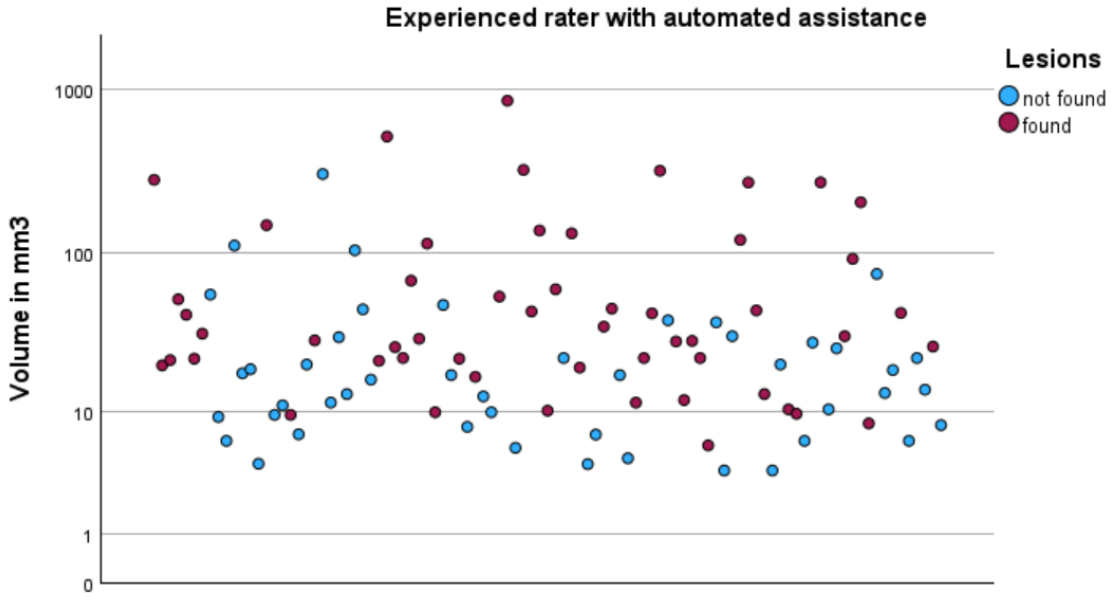

Figure 6: Detection of new lesions by the experienced rater with automated assistance. The y-axis represents lesion volume in mm^3 on a logarithmic scale. Lesions detected by the rater are shown in red ("found"), while those missed are shown in blue ("not found").

Without using the AI-based software, the lesions correctly identified by the experienced rater had an average size of 93.06 mm^3 (95% CI: 49.32 – 136.81 mm^3), while with automated assistance, the average lesion size identified was 91.57 mm^3 (95% CI: 50.49 – 132.65 mm^3). Although lesions identified with automation were on average smaller, this size difference was not statistically significant due to the considerable overlap of the 95% CIs. The same applied to the undetected size progressive lesions: the average volume of these lesions without software assistance was 32.15 mm^3 (95% CI: 16.38 – 47.92 mm^3) and with assistance 28.58 mm^3 (95% CI: 14.35 – 42.80 mm^3). Due to the overlapping 95% CIs, no significant difference could be determined. With or without software

assistance, the lesions identified as progressive were significantly larger than those that were not detected.

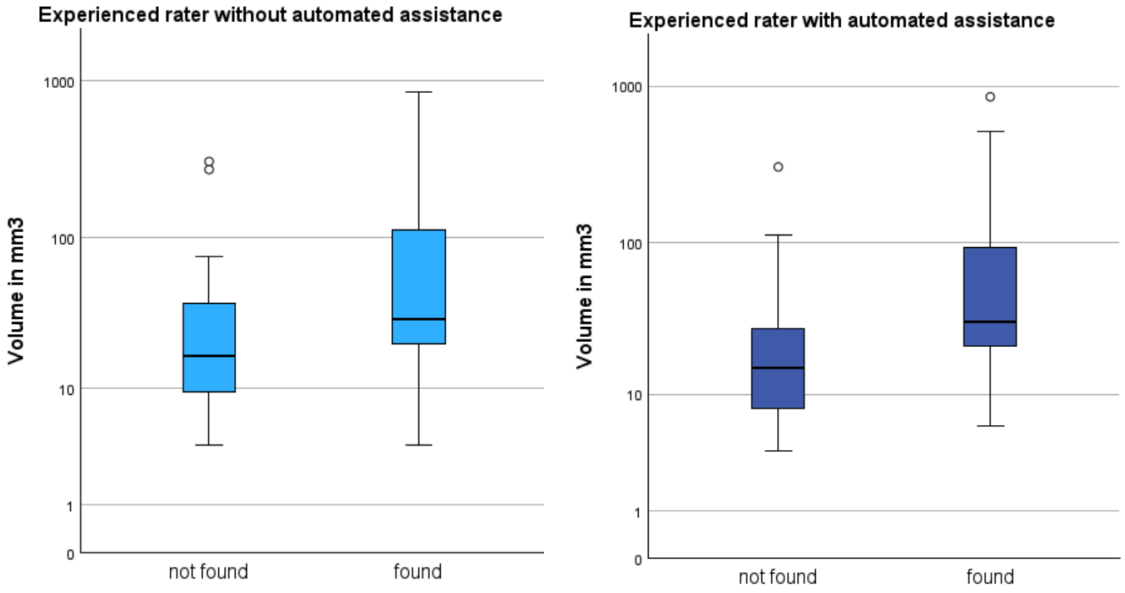

Figure 7: Boxplots showing lesion volumes (in mm³) for new lesions missed ("not found") and detected ("found") by the experienced rater, without (left) and with (right) automated assistance. Volumes are displayed on a logarithmic scale.

4.3.1.2 Results of the inexperienced rater

Without support of the AI-based software, the inexperienced rater classified 79 out of 99 new lesions as progressive with a sensitivity of 0.80, and the other 20 lesions as non-progressive. The statistical analysis showed the following results: the lesions identified had an average size of 73.11 mm³ (95% CI: 44.13 – 102.09 mm³) with an SD of 129.4 mm³, while the lesions incorrectly classified as negative had an average volume of 19.60 mm³ (95% CI: 12.57 – 26.63 mm³) with an SD of 15.02 mm³. This resulted in an average volume difference of 53.51 mm³. An independent t-test ($t(97) = 1.839$, $p = 0.069$) indicated a trend toward smaller lesions being more likely to be missed. However, this difference did not reach statistical significance, although the non-overlapping 95% CIs suggested a significant volume difference between lesions correctly identified as progressive and lesions incorrectly classified as non-progressive.

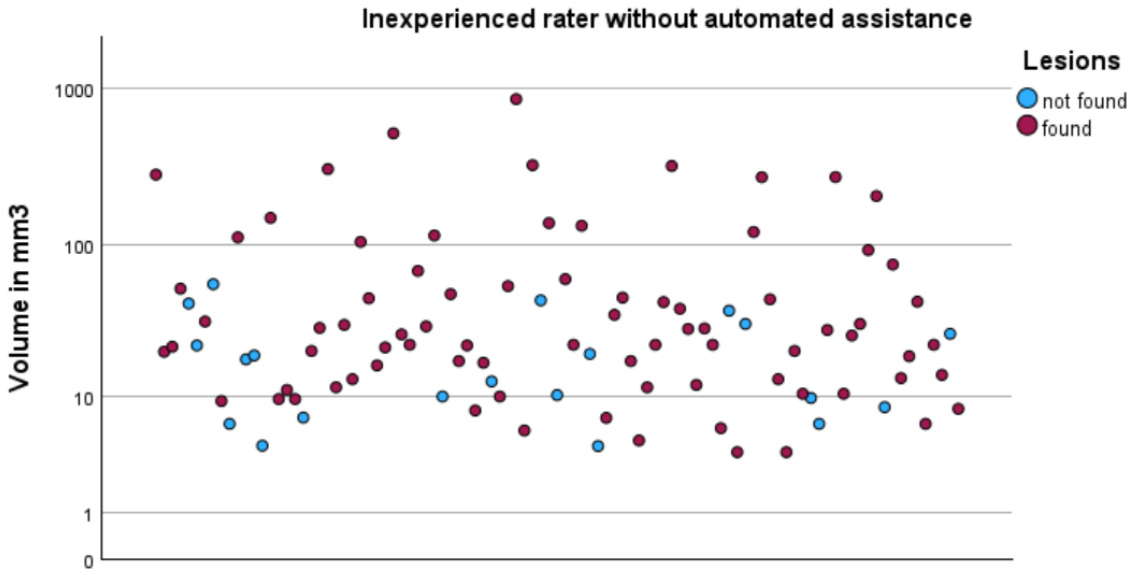

Figure 8: Detection of new lesions by the inexperienced rater without automated assistance. The y-axis shows lesion volume in mm³ on a logarithmic scale. Lesions identified by the rater are marked in red ("found"), while missed lesions are shown in blue ("not found").

With automated assistance, 87 of the new lesions were correctly identified as progressive, and 12 were incorrectly identified as non-progressive or not recognized as new. In this case, the sensitivity was 0.88. The statistical analysis showed that the identified lesions had an average size of 66.30 mm^3 (95% CI: 39.75 – 92.85 mm^3) with an SD of 124.57 mm^3, while the lesions classified as non-progressive had an average size of 33.32 mm^3 (95% CI: 14.00 – 52.64 mm^3) with an SD of 30.41 mm^3. In this case, the volume difference between the detected and undetected new lesions was 32.98 mm^3. An independent t-test ($t(97) = 0.91$. $p = 0.365$) indicated that the observed difference in volume between FN lesions and TP lesions was not statistically significant. This is also supported by the fact that the 95% CIs do overlap.

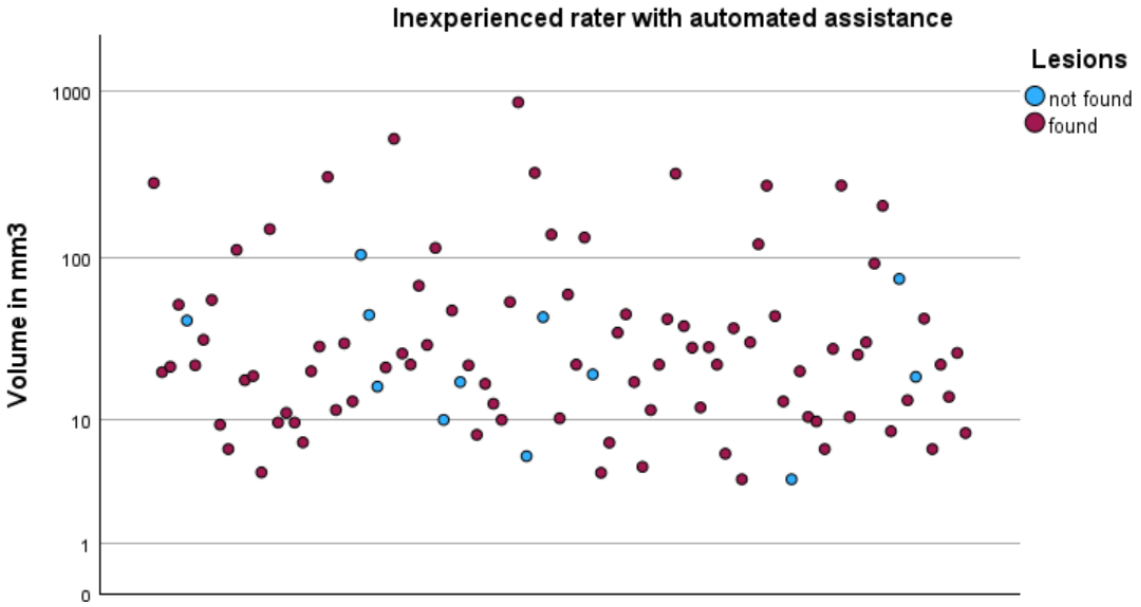

Figure 9: Detection of new lesions by the inexperienced rater with automated assistance. The y-axis represents lesion volume in mm³ on a logarithmic scale. Lesions detected by the rater are marked in red ("found"), while missed lesions are shown in blue ("not found").

Without the aid of the AI-based software, lesions which were correctly identified by the inexperienced rater had an average size of 73.11 mm^3 (95% CI: 44.13 – 102.09 mm^3), while those identified with automated assistance had an average size of 66.3 mm^3 (95% CI: 39.75 – 92.85 mm^3). The overlapping 95% CIs suggested that despite the average size of lesions identified with automated assistance being smaller, this difference may not be statistically significant. Without the automated program, the lesions incorrectly identified as non-progressive had an average volume of 19.6 mm^3 (95% CI: 12.57 – 26.63 mm^3). With automated assistance, the FN lesions had an average size of 33.32 mm^3 (95% CI: 14.00 – 52.64 mm^3). The overlap in 95% CIs suggested that the difference in average size of the undetected lesions with and without automated assistance may not be statistically significant.

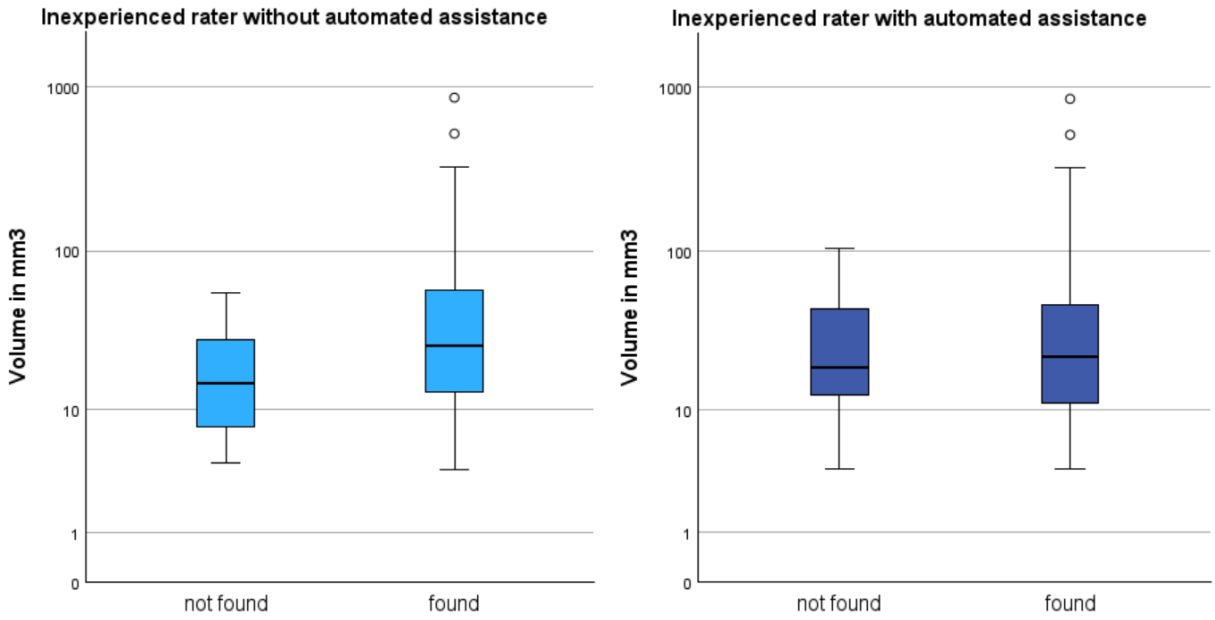

Figure 10: Boxplots showing lesion volumes in mm³ for new lesions missed ("not found") and detected ("found") by the inexperienced rater on a logarithmic scale. The left figure shows the volumes without automated assistance, while the right figure shows the volumes with automated assistance.

4.3.1.3 Results of the AI-based program

Out of the 99 new lesions according to the ground truth, the AI-based software identified 28 as progressive, while the remaining 71 new lesions were classified as non-progressive. This resulted in a sensitivity of 0.28. The TP lesions had an average volume of 149.09 mm^3 (95% CI: 78.19 – 219.99 mm^3) with an SD of 182.85 mm^3. For the lesions not identified as progressive, the average volume was 28.07 mm^3 (95% CI: 16.78 – 39.36 mm^3) with an SD of 47.7 mm^3. The difference in the average volumes of detected and undetected lesions in this case was 121.02 mm^3. An independent t-test ($t(97) = 5.183$, $p < 0.0001$) showed that the difference in volume was statistically significant, which is further supported by the fact that the 95% CIs for detected versus undetected lesions did not overlap.

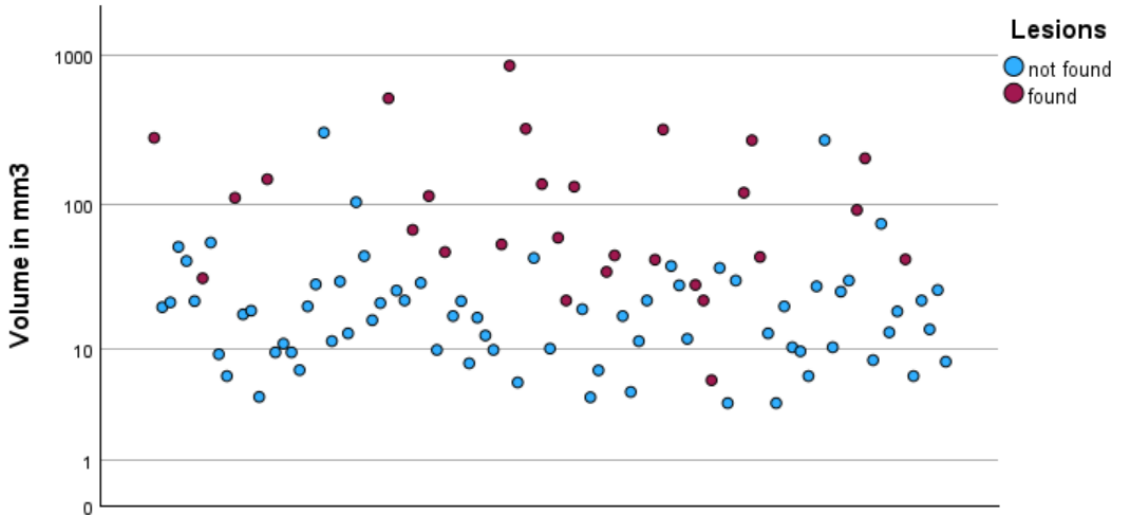

Figure 11: Detection of new lesions by the AI-based program. The y-axis shows lesion volume in mm³ on a logarithmic scale. Lesions identified are marked in red ("found"), while missed lesions are shown in blue ("not found").

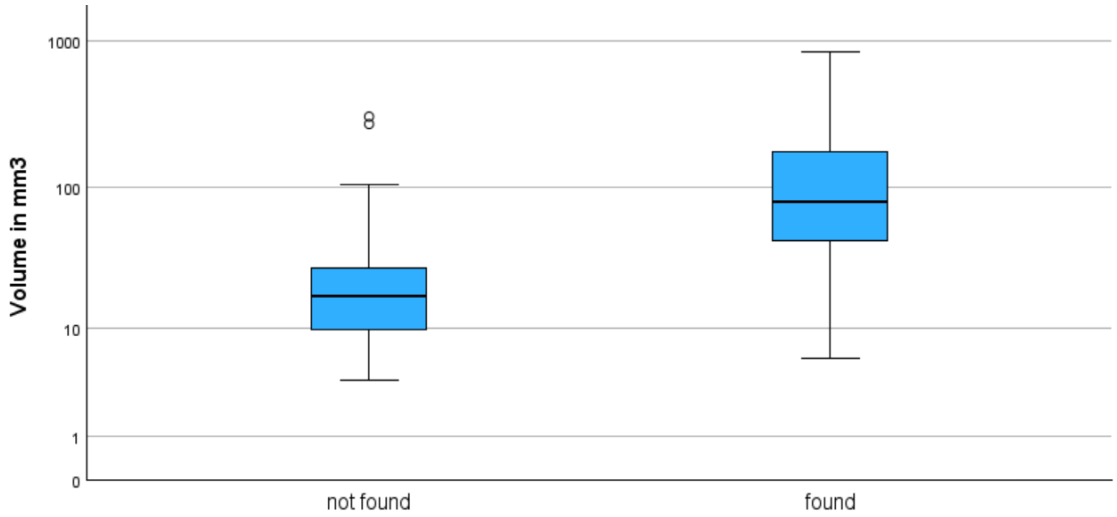

Figure 12: Boxplots comparing lesion volumes (in mm³) for missed ("not found") and detected ("found") new lesions by the AI-based program. Lesion volumes are displayed on a logarithmic scale.

4.3.1.4 Comparison of statistical parameters: experienced vs. inexperienced rater

Although the inexperienced rater identified smaller lesions on average both without (mean volume 73.11 mm³, 95% CI 44.13 – 102.09 mm³) and with automated assistance (mean volume 66.3 mm³, 95% CI 39.75 – 92.85 mm³) compared to the experienced rater (without assistance: mean volume 93.06 mm³,

95% CI 49.32 – 136.81 mm^3; with assistance: mean volume 91.57 mm^3, 95% CI 50.49 – 132.65 mm^3), the overlapping 95% CIs suggested that this volume difference was not statistically significant. Similarly, no statistically significant difference in volume was found between the experienced (without assistance: mean volume 32.15 mm^3, 95% CI 16.38 – 47.92 mm^3; with assistance: mean volume 28.58 mm^3, 95% CI 14.35 – 42.80 mm^3) and inexperienced rater (without assistance: mean volume 19.6 mm^3, 95% CI 12.57 – 26.63 mm^3; with assistance: mean volume 33.32 mm^3, 95% CI 14.00 – 52.64 mm^3) for the undetected lesions, regardless of whether assistance was used. For both the experienced and inexperienced rater, the size of the newly detected lesions did not differ significantly with use of the automated program compared to the assessment without it. The same applied to the undetected new lesions. For the experienced rater, the detected lesions were significantly larger than the undetected lesions, both with and without software assistance. For the inexperienced rater, this was only true for the assessment without automated assistance, whereas with software support, no significant volume difference was observed between detected and undetected lesions. For the AI-based tool alone, the detected lesions (mean volume 149.09 mm^3, 95% CI 78.19 – 219.99 mm^3) were significantly larger than the undetected lesions (mean volume 28.07 mm^3, 95% CI 16.78 – 39.36 mm^3). The volume of the detected lesions exceeded that of the lesions identified by either the experienced or inexperienced rater. However, the overlapping 95% CIs in both cases indicated no significant difference in volume between the lesions detected and undetected by the automated tool compared to those by either rater.

4.3.2 *Volume of enlarged lesions*

Out of the 122 progressive lesions, 23 were enlarged lesions. In addition to measuring the volume of the lesions at the second MRI time point, the volume difference between the two MRI scans was recorded as well. The following section describes the results of the lesions identified as progressive and non-progressive, based on the volume differences for each rater. Since the data for

enlarged lesions was not normally distributed, the median and interquartile range were used to describe it.

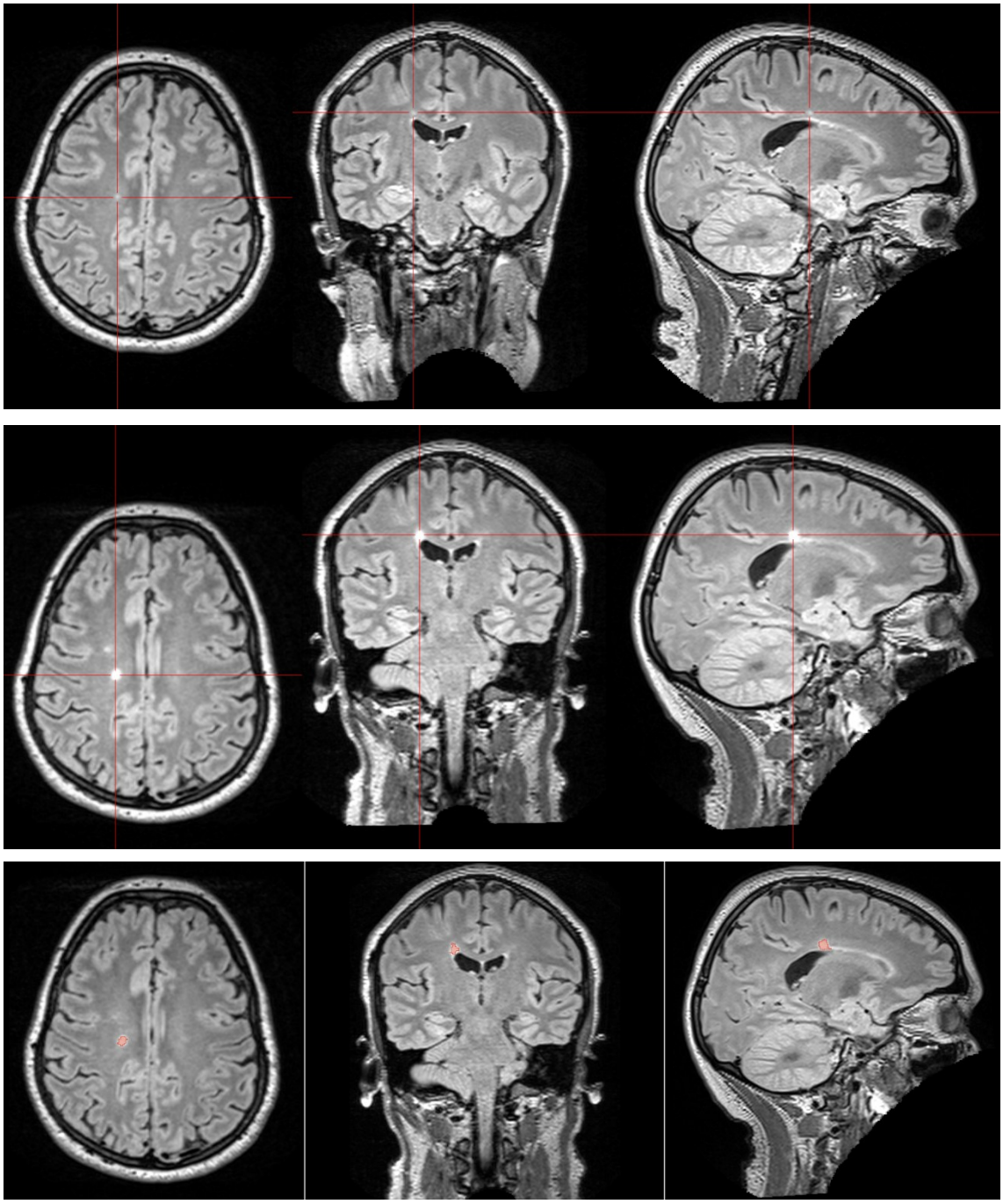

Figure 13: Example of an enlarged MS lesion. The first image shows timepoint 1, the second shows timepoint 2, and the third displays timepoint 2 with color-coded segmentation of the enlarged lesion using 3D Slicer.

4.3.2.1 Results of the experienced rater

Of the 23 enlarged lesions, the experienced rater accurately classified 10 as progressive without automated assistance, resulting in a sensitivity of 0.43. The volume increase of the lesions was used to statistically compare these with the lesions not identified as progressive. The statistical analysis showed that the detected progressive lesions had an average volume increase of 192.82 mm^3 (95% CI: 42.02 – 343.61 mm^3) compared to the initial MRI examination, with a median of 133.09 mm^3. The first quartile was 37.46 mm^3 and the third quartile was 258.28 mm^3, resulting in an interquartile range of 220.82 mm^3. The remaining 13 progressively enlarged lesions were not detected and were therefore considered FN. For these undetected lesions, the analysis revealed an average volume increase of 105.73 mm^3 (95% CI: 9.50 – 201.96 mm^3) with a median of 49.95 mm^3. The first quartile was 20.38 mm^3 while the third quartile was 104.50 mm^3, giving an interquartile range of 84.12 mm^3. Since the values were not normally distributed, a Wilcoxon test was performed to determine whether the lesion volumes differed significantly based on medians. The Wilcoxon test produced a p-value of 0.17, indicating no statistically significant difference in median volumes of detected and undetected enlarged lesions. Although lesions with larger volume increases were more likely to be identified as progressive compared to those with smaller increases, this difference was not statistically significant based on median volumes.

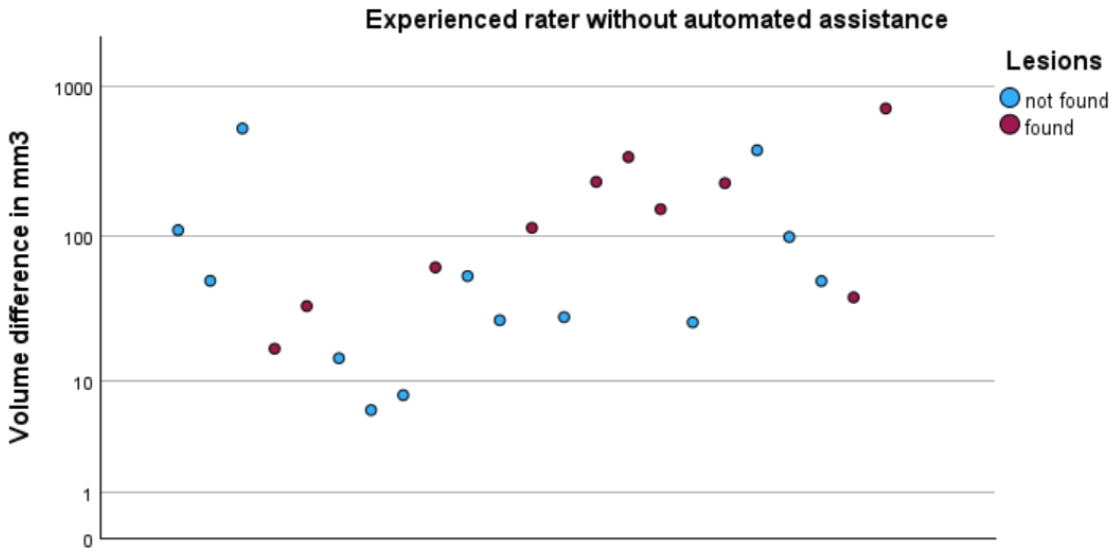

Figure 14: Identification of enlarged lesions by the experienced rater without automated support. The y-axis displays the lesion's volume difference between timepoint 1 and timepoint 2 in mm^3 on a logarithmic scale. Lesions recognized by the rater are shown in red ("found"), while those missed are marked in blue ("not found").

With assistance from the automated program, the experienced rater correctly classified 12 of the 23 enlarged progressive lesions as progressive, while misclassifying 11 as non-progressive, which resulted in a sensitivity of 0.52. The volume increase between the two MRI imaging points was recorded, and a Wilcoxon test was performed to compare the median volumes of detected and undetected lesions. For the 12 lesions identified as progressive, the average volume increase was 177.05 mm^3 (95% CI: 50.59 – 303.51 mm^3) with a median of 106.54 mm^3. The first quartile was 50.96 mm^3 and the third quartile 230.39 mm^3, resulting in an interquartile range of 179.43 mm^3. For the 11 lesions not classified as progressive, the average volume increase was 107.1 mm^3 (95% CI: -5.96 – 220.16 mm^3), with a median of 33.75 mm^3. The first quartile for these lesions was 14.65 mm^3 and the third quartile was 110.00 mm^3, yielding an interquartile range of 95.35 mm^3. The Wilcoxon test produced a p-value of 0.18, which indicated that the difference in volume increase between detected and undetected lesions was not statistically significant, although the experienced rater was generally more likely to detect progressive lesions with larger volume increases.

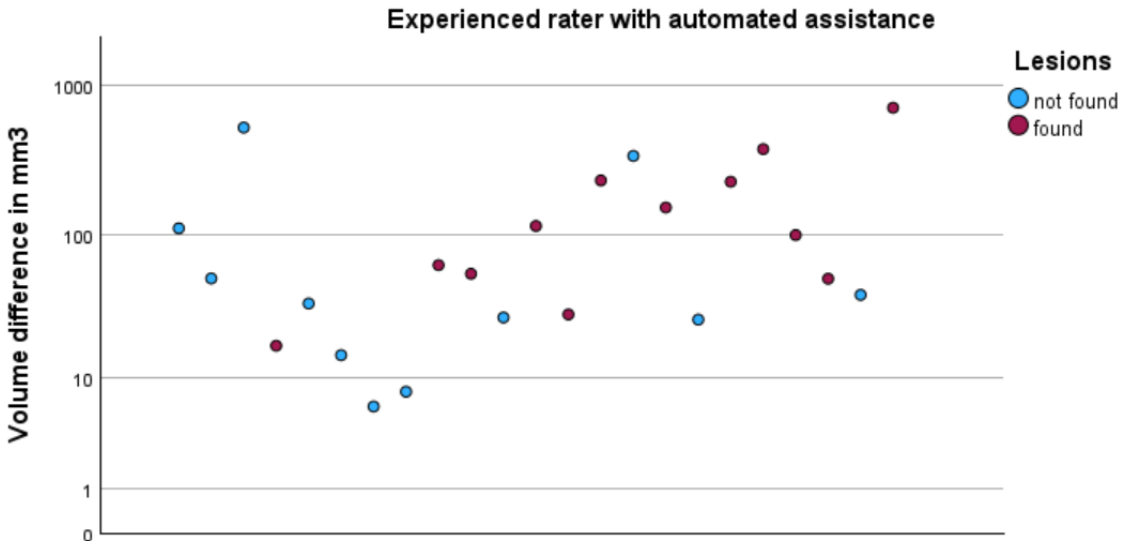

Figure 15: Detection of enlarged lesions by the experienced rater with automated assistance. The y-axis shows the lesion's volume difference between timepoint 1 and timepoint 2 in mm^3 on a logarithmic scale. Lesions identified by the rater are marked in red ("found"), while missed lesions are shown in blue ("not found").

Without support from the automated software, the lesions identified as enlarged by the experienced rater had a median volume increase of 133.09 mm^3. With the support of the automated program, the rater identified lesions with a smaller median volume increase of 106.54 mm^3. The statistical analysis using a Wilcoxon test yielded a p-value of 0.46, indicating that this difference was not statistically significant. The interquartile range of 220.82 mm^3 without automated support indicated a broader spread compared to 179.43 mm^3 with software support. For lesions classified as non-progressive, the median volume increase was 49.95 mm^3 without software assistance and 33.75 mm^3 with assistance. Here too, the Wilcoxon test yielded no significant difference, with a p-value of 0.91, though the experienced rater tended to classify lesions with a larger median volume increase as non-progressive when not using AI-based assistance. For undetected lesions, data variability was higher with software support, showing an interquartile range of 95.35 mm^3 compared to 84.12 mm^3 without assistance. Although the experienced rater tended to classify lesions with a larger volume increase as enlarged both with and without automated support, the statistical analysis in both cases showed that these differences were not significant.

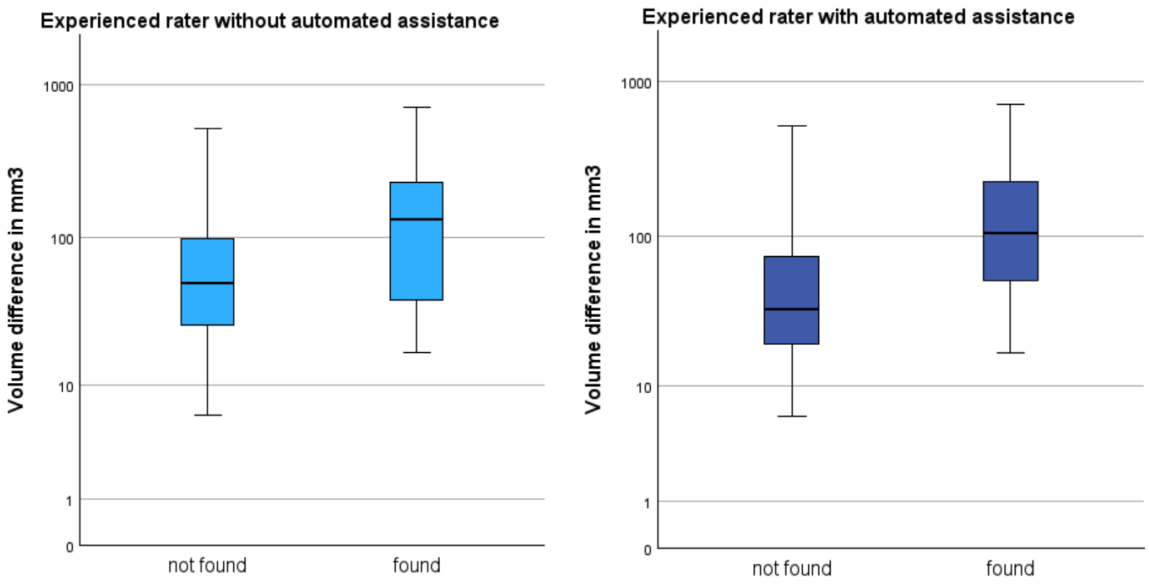

Figure 16: Boxplots showing the volume differences between timepoint 1 and timepoint 2 (in mm^3) for enlarged lesions that were missed ("not found") or detected ("found") by the experienced rater, without (left) and with (right) automated assistance. The volume differences are displayed on a logarithmic scale.

4.3.2.2 Results of the inexperienced rater

Out of the 23 enlarged progressive lesions, the inexperienced rater classified 16 as progressive without the AI-based software, while 7 lesions were not detected. This corresponded to a sensitivity of 0.70. The statistical analysis showed that the average volume difference for the 16 lesions classified as enlarged was 168.57 mm^3 (95% CI: 58.95 – 278.2 mm^3), with a median of 85.94 mm^3. The interquartile range was 203.05 mm^3, with a first quartile of 27.34 mm^3 and a third quartile of 230.39 mm^3. For the seven lesions not classified as progressive, the average volume difference was 86.5 mm^3 (95% CI: -19.87 – 192.86 mm^3), with a median of 50.18 mm^3. The interquartile range for these lesions was 81.90 mm^3, with a first quartile of 17.10 mm^3 and a third quartile of 99.00 mm^3. While the inexperienced rater was generally more likely to correctly identify lesions with a larger volume increase as progressive, the Wilcoxon test performed showed a p-value of 1.00, suggesting that the difference in median volume increase between detected and undetected lesions was not statistically significant.

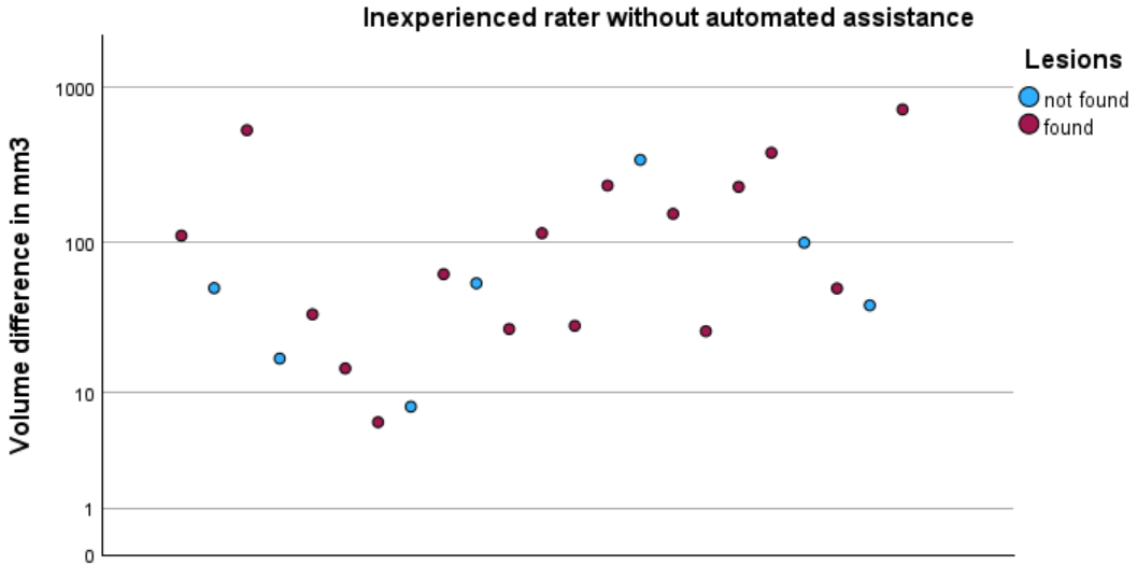

Figure 17: Identification of enlarged lesions by the inexperienced rater without automated support. The y-axis shows the lesion's volume difference between timepoint 1 and timepoint 2 in mm^3 on a logarithmic scale. Lesions identified by the rater are marked in red ("found"), while missed lesions are shown in blue ("not found").

With automated assistance, the inexperienced rater correctly classified 17 of the 23 enlarged progressive lesions as enlarged, yielding a sensitivity of 0.74. The average volume difference between the two MRI time points for the 17 lesions classified as enlarged was 161.73 mm^3 (95% CI: 58.38 – 265.09 mm^3), with a median of 61.87 mm^3. The first quartile was 27.68 mm^3 and the third quartile was 229.26 mm^3, resulting in an interquartile range of 201.58 mm^3. For the remaining 6 lesions that were not classified as progressive, the average volume difference was 92.21 mm^3 (95% CI: -39.79 – 224.20 mm^3), with a median of 36.23 mm^3. The interquartile range for these lesions was 146.05 mm^3, with a first quartile of 21.10 mm^3 and a third quartile of 167.15 mm^3. The Wilcoxon test yielded a p-value of 0.92, indicating that the volume difference between detected and undetected lesions was not statistically significant.

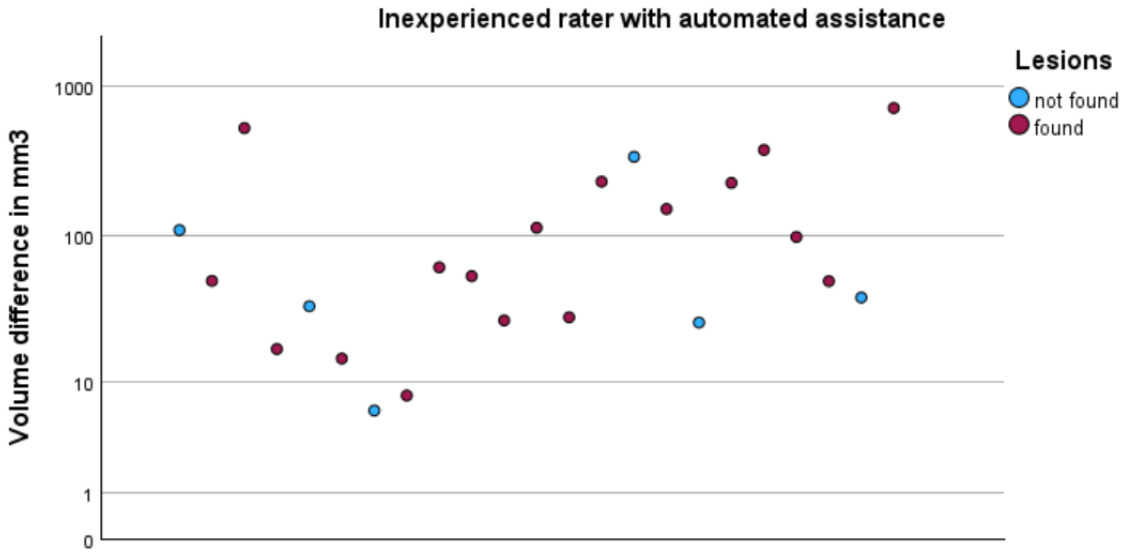

Figure 18: Detection of enlarged lesions by the inexperienced rater with automated assistance. The y-axis displays the lesion's volume difference between timepoint 1 and timepoint 2 in mm^3 using a logarithmic scale. Lesions identified by the rater are marked in red ("found"), while undetected lesions are shown in blue ("not found").

Without the AI-based software, lesions classified as progressive had an average median volume of 85.94 mm^3. In contrast, the inexperienced rater using the software identified lesions as progressive with a median volume increase of 61.86 mm^3. The Wilcoxon test showed no statistically significant difference, with a p-value of 0.79. The spread of values, measured by interquartile range, was similar without (203.05 mm^3) and with (201.58 mm^3) automated support. Lesions that were incorrectly not identified as progressive had an average median volume increase of 50.18 mm^3 without software assistance, compared to 36.23 mm^3 with automated support. Here, the Wilcoxon test yielded a p-value of 0.92, again indicating no statistically significant difference, though the inexperienced rater with AI support identified lesions with a smaller median volume increase as progressive. Without software support, the spread of values was lower (81.90 mm^3) compared to the spread with software support (146.05 mm^3). Although the inexperienced rater tended to classify lesions with larger volume increases as progressive both with and without automated support, statistical analysis in both cases found these differences to be insignificant.

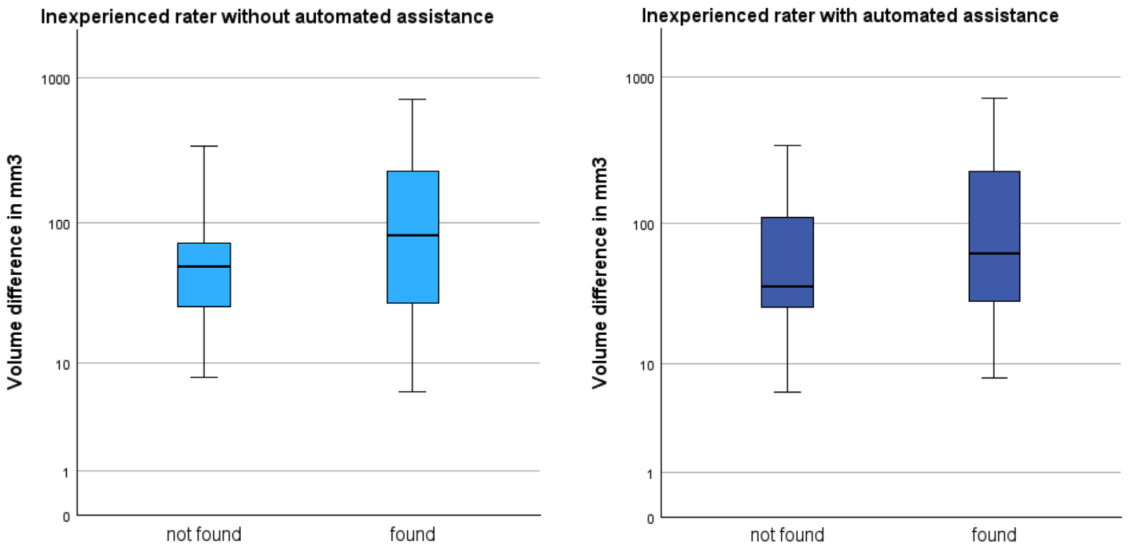

Figure 19: Boxplots illustrating the volume differences between timepoint 1 and timepoint 2 (in mm³) of enlarged lesions either missed ("not found") or identified ("found") by the inexperienced rater, presented without (left) and with (right) automated assistance. The volume differences are shown on a logarithmic scale.

4.3.2.3 Results of the AI-based program

Out of the 23 size progressive lesions identified by the ground truth, the automated assessment software correctly classified 15 as progressive, yielding a sensitivity of 0.65. The average volume increase for these 15 detected lesions between the two MRI time points was 201.2 mm^3 (95% CI: 86.86–315.55 mm^3), with a median increase of 114.07 mm^3. The interquartile range was 288.42 mm^3, with the first quartile at 50.18 mm^3 and the third quartile at 338.60 mm^3. For the eight lesions not identified as enlarged, the average volume increase between the imaging time points was 35.58 mm^3 (95% CI: 9.00–62.16 mm^3), with a median of 27.68 mm^3. Here, the interquartile range was 19.95 mm^3, with the first quartile at 17.51 mm^3 and the third quartile at 37.46 mm^3. Although the software was more likely to classify lesions with larger volume increases as progressive, the Wilcoxon test yielded a p-value of 0.07, indicating that the difference in volume increase between detected and undetected lesions was not statistically significant.

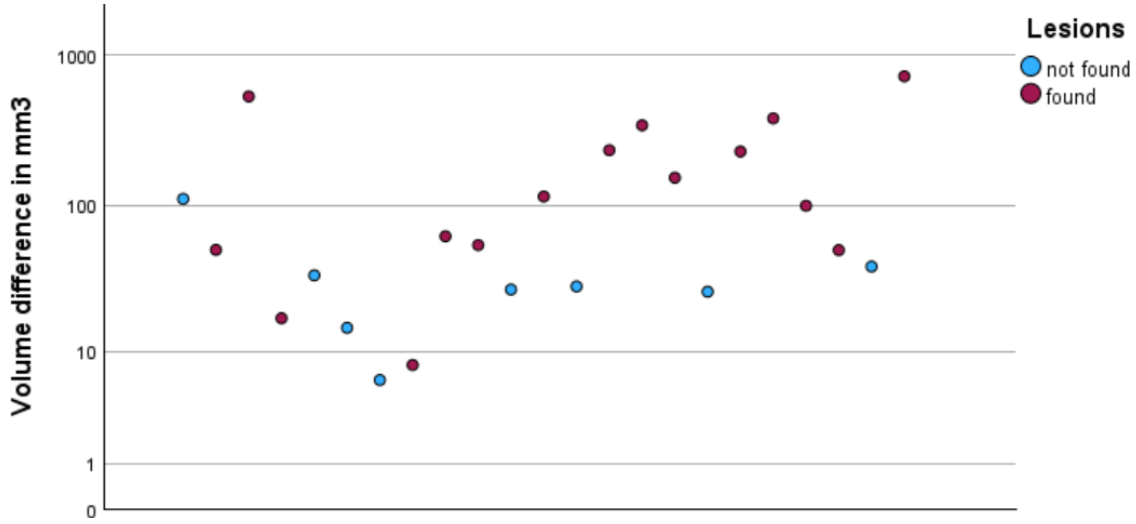

Figure 20: Identification of enlarged lesions by the automated program. The y-axis displays the lesion's volume difference between timepoint 1 and timepoint 2 in mm³ using a logarithmic scale. Lesions identified by the AI-based program are marked in red ("found"), while undetected lesions are shown in blue ("not found").

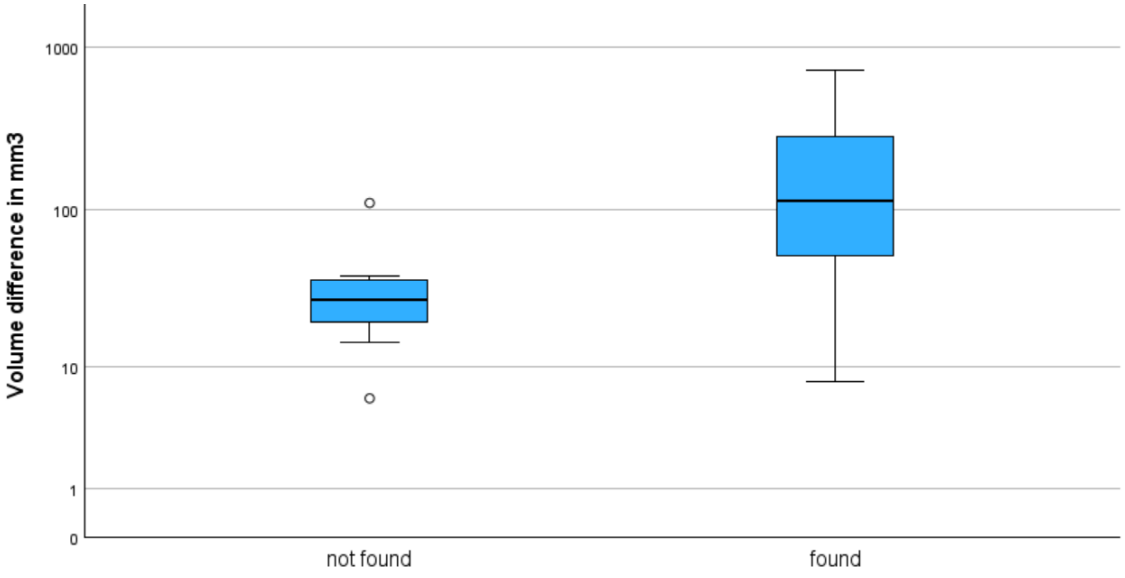

Figure 21: Boxplots displaying the volume differences between timepoint 1 and timepoint 2 (in mm³) of enlarged lesions that were missed ("not found") or identified ("found") by the automated program, shown on a logarithmic scale.

4.3.2.4 Comparison of statistical parameters: experienced vs. inexperienced rater

The assessment of enlarged lesions by the experienced rater, the inexperienced rater and the automated program highlighted clear differences in the median

volume increases for lesions identified as progressive and those missed. Across all raters, lesions classified as progressive had larger median volume increases than those undetected, though none of the approaches showed a statistically significant difference in these volume changes, as reflected by p-values from Wilcoxon tests in the statistical analysis.

Without the AI-based software, the experienced rater identified progressive lesions with a median volume increase of 133.09 mm^3 compared to the initial examination. In contrast, the inexperienced rater classified lesions as progressive at a lower median volume increase of 85.94 mm^3, indicating that the inexperienced rater detected lesions with smaller volume increases compared to the experienced rater. To assess if this difference was statistically significant, a Wilcoxon test was applied, resulting in a p-value of 0.14, which did not indicate statistical significance. For lesions missed without automated aid, the median volume increase was 49.95 mm^3 for the experienced rater and 50.18 mm^3 for the inexperienced rater. Here, the Wilcoxon test yielded a p-value of 1.00, showing no significant difference between the raters.

With automated support, lesions identified as progressive by the experienced rater had a median volume increase of 106.51 mm^3, while the inexperienced rater identified progressive lesions at an even lower median volume increase of 61.87 mm^3. However, the Wilcoxon test yielded a p-value of 0.16, showing no statistically significant difference between the raters in this round either. Among the undetected lesions, the experienced rater missed lesions with a median volume increase of 33.75 mm^3, which was lower than for the inexperienced rater (36.23 mm^3), yet the Wilcoxon test resulted in a p-value of 0.89, again indicating no statistically significant difference.

Both with and without AI-based assistance, there was no statistically significant difference between the experienced and inexperienced raters in their classification of progressive lesions, despite the inexperienced rater generally identifying lesions with smaller median volume increases.

4.4 Evaluation duration

With the growing number of radiological examinations (European Society of Radiology 2009 2010), the efficiency and duration of evaluating MRI images for MS patients are crucial in clinical routine. Therefore, in researching this automated tool for assisting in the assessment of MRI images, the question of the difference in duration with versus without the automated program is an important parameter.

4.4.1 *Evaluation duration of the experienced rater*

For the experienced raters, only the total assessment duration, yet not the time for individual cases, was recorded. Therefore, no statistical analysis of the experienced rater's values could be conducted.

The 110 patient cases were split into two rounds, with one experienced rater evaluating each round. The first round consisted of 53 cases, which were evaluated by a neuroradiologist with 16 years of experience. Without automated support, the rater took an average of 03:47 minutes per case, while with software assistance, this was reduced to 02:02 minutes. With the software support, the rater saved 01:45 minutes per case, representing a percentage reduction of 46.32%.

The remaining 57 cases were evaluated in round 2 by a neuroradiologist with 8 years of experience. Without software support, the evaluation took an average of 06:37 minutes per case, whereas with assistance, the time was reduced to 04:21 minutes. This means the rater was 02:16 minutes faster per case when using the AI-based program, which corresponds to a percentage reduction of 34.23%.

When combining both rounds, the experienced raters took an average of 05:15 minutes per case to evaluate all 110 cases without the assistance of the automated program, whereas with automated support, the time was reduced to 03:14 minutes. This represents a difference of 02:01 minutes or a percentual reduction of 38.41%.

4.4.2 *Evaluation duration of the inexperienced rater*

Since the inexperienced rater had no prior experience with radiology, a learning effect during the initial evaluation of MRI images of MS patients could not be ruled out. Therefore, an initial assessment of the duration per case was conducted using a paired bar chart, assuming that outliers with particularly long evaluation times would appear in the beginning. The paired bar chart showed a noticeably large difference only for the first evaluated case. This was most likely due to the learning effect, so the statistical analysis of the time required for the evaluation was calculated both with and without this case.

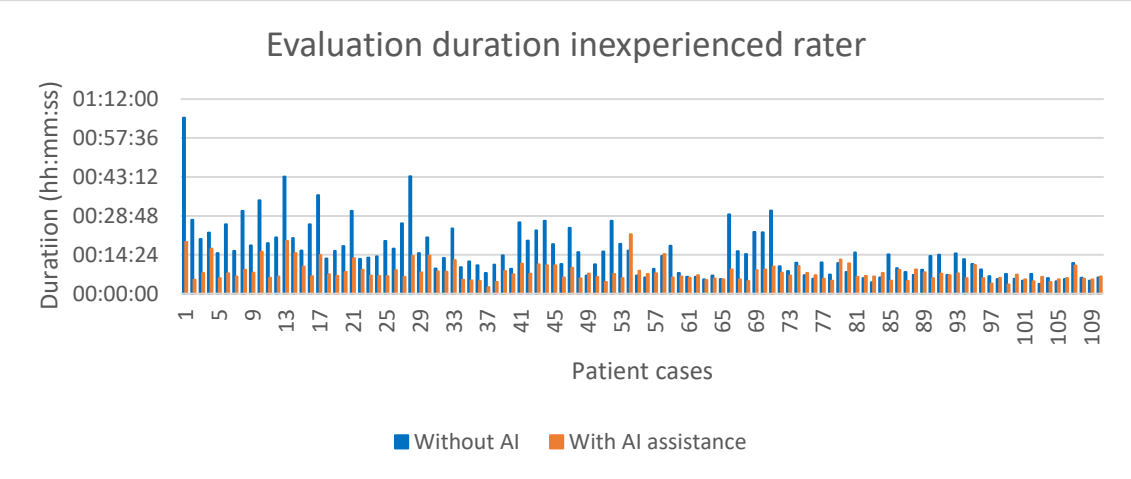

Figure 22: Evaluation duration per case for all 110 cases assessed by the inexperienced rater, shown with (orange) and without (blue) AI assistance. Each bar represents the time required to evaluate a single case.

4.4.2.1 Without automated assistance

Without the assistance of the automated program, the inexperienced rater required an average of 14:59 minutes per case (95% CI: 13:07 min – 16:51 min) for the evaluation of both rounds. The SD was 09:50 minutes, the shortest duration for the assessment of one case was 03:41 minutes, while the longest was 01:05:07 hours.

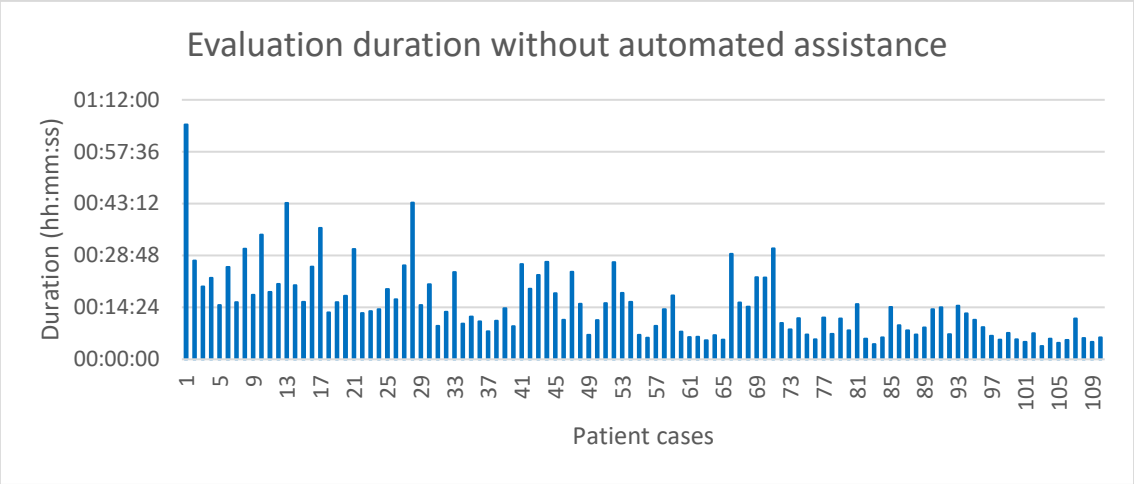

Figure 23: Evaluation duration per case for all 110 cases assessed by the inexperienced rater without AI assistance. Each bar represents the time taken to evaluate a single case.

Additionally, a statistical analysis was conducted excluding the first case, which was considered an outlier due to its significantly longer evaluation time of 01:05:07 hours, attributed to the learning effect. The evaluation of the remaining 109 cases resulted in an average duration of 14:31 minutes per case (95% CI: 12:53 min – 16:10 min) with an SD of 08:37 minutes.

Furthermore, the evaluation process for the inexperienced rater was divided into two rounds (the first with 53 cases and the second with 57 cases) to assess a possible learning effect over the course of a prolonged observation. In the following analysis, both the duration for both rounds and the duration of each round individually were evaluated.

In the first evaluation round, 53 cases were assessed with an average duration of 20:22 minutes per case (95% CI: 17:29 min – 23:15 min) with an SD of 10:27 minutes. The first case was considered an outlier due to the learning effect. Excluding this outlier, the statistical analysis for the first round showed an average duration of 19:30 minutes (95% CI: 17:09 min – 21:51 min) with an SD of 08:26 minutes.

In the second evaluation round, 57 cases were assessed with an average duration of 09:59 minutes (95% CI: 08:25 min – 11:32 min) per case and an SD of 05:52 minutes.

Since the 95% CIs for the time per case in the first and second round did not overlap, it can be assumed that the inexperienced rater had significantly shorter evaluation times per case in the second round.

Additionally, a Pearson correlation was calculated to determine the strength of relationship between the evaluation duration of the two rounds. This resulted in a correlation of 0.30 ($p < 0.05$), indicating only a weak association.

4.4.2.2 With automated assistance

With assistance of the automated program, the inexperienced rater required an average of 07:56 minutes per case (95% CI: 07:17 min – 08:35 min) for the assessment of all 110 cases. The SD was 03:27 minutes, the shortest duration for the evaluation of one case was 02:27 minutes and the longest was 21:58 minutes. The times for all cases were displayed in a bar chart.

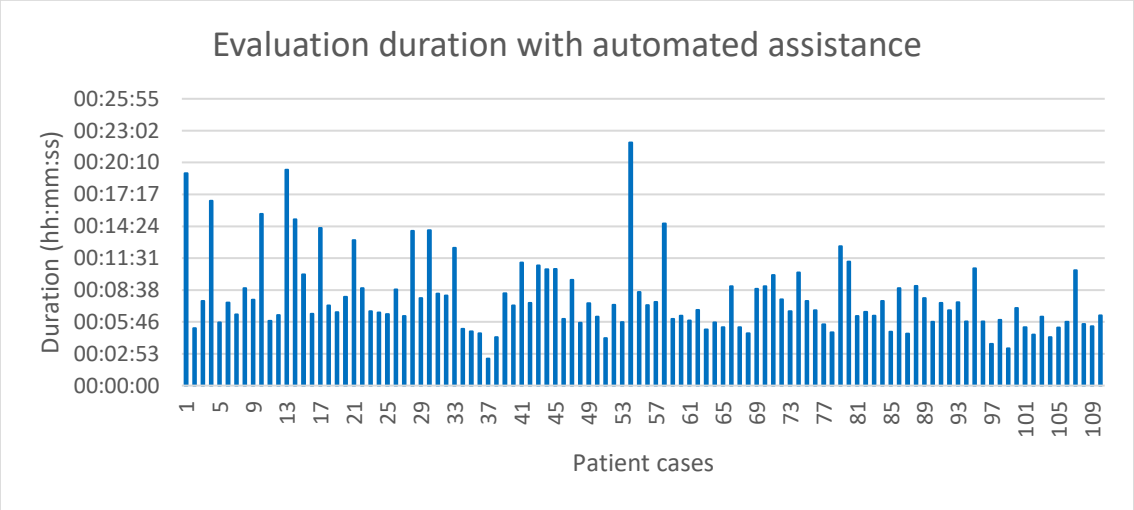

Figure 24: Time taken to evaluate each of the 110 cases by the inexperienced rater using AI assistance. Each bar shows the duration spent on one individual case.

The first round of evaluations involved MRI data from 53 MS patients. With the help of the AI-based software, the inexperienced rater took an average of 08:42 minutes per case (95% CI: 07:39 min – 09:45 min), with an SD of 03:48 minutes. For the evaluation of the other 57 cases in the second round, the average time per case was 07:13 minutes (95% CI: 06:26 min – 08:00 min), with an SD of 02:57 minutes.

In the second round, the inexperienced rater took less time per case. However, the 95% CIs for the first and second round did overlap, so this difference was not statistically significant.

A Pearson correlation was calculated to assess the strength of relationship between the evaluation durations of the first and second round. The calculation yielded a Pearson correlation of 0.29 ($p < 0.05$), indicating only a weak association between the two rounds.

4.4.2.3 Comparison of evaluation duration without vs. with automatic assistance

For the statistical analysis of evaluation duration without versus with automatic assistance, a paired t-test was conducted. The analysis showed a significant difference of 07:03 minutes (95% CI: 05:30 – 08:35 min) in evaluation duration without the automated program (mean (M) = 14:59 min, SD = 09:50 min) compared to with the program (M = 07:56 min, SD = 03:27 min), $t(109) = 9.04$, $p < 0.001$. The effect size, measured by Cohen's d, was $d = 0.86$ (95% CI: 0.64 – 1.08), indicating a strong effect.

Additionally, an analysis was conducted excluding the first case as it was classified as an outlier, since its unusually long evaluation time might have been due to a learning effect. Without this outlier, the paired-samples t-test showed a significant difference of 06:41 minutes (95% CI: 05:18 – 08:04 min) in evaluation time without AI-based support (M = 14:31 min, SD = 08:37 min) compared to with support (M = 07:50 min, SD = 03:17 min), $t(108) = 9.56$, $p < 0.001$. The effect size, calculated with Cohen's d, was $d = 0.92$ (95% CI: 0.69 – 1.14), indicating a strong effect.

Furthermore, a Pearson correlation coefficient was calculated to compare the duration without and with automated assistance, which showed a strong correlation of 0.62 ($p < 0.001$). Even when excluding the outlier, a Pearson correlation coefficient of 0.56 ($p < 0.001$) still indicated a large effect.

In addition to analyzing all patient cases, a comparison of evaluation duration with and without automated support was also conducted separately for the first round with 53 cases and the second round with 57 cases.

The paired t-test for the statistical analysis of the first 53 cases revealed a statistically significant difference of 11:40 minutes (95% CI: 9:26 min – 13:53 min) in evaluation duration without (M = 20.22 min, SD = 10:27 min) versus with automated assistance (M = 08:42 min, SD = 03:48 min), $t(52) = 10.53$, $p < 0.001$. The effect size using Cohen's d showed a strong intervention effect with $d = 1.45$ (95% CI: 1.06 – 1.83).

When excluding the outlier in the paired t-test, there was a statistically significant difference in duration of 11:00 minutes (95% CI: 09:11 min – 12:50 min) between the evaluation without (M = 19:30 min, SD = 08:26 min) and with the AI-based program (M = 08:30 min, SD = 03:32 min), $t(51) = 12.11$, $p < 0.001$. Here, the effect size using Cohen's d indicated a large effect with $d = 1.68$ (95% CI: 1.25 – 2.10) as well.

For the first 53 cases, Pearson correlation analysis showed a strong relationship between processing times with and without automated support of 0.74 ($p < 0.001$). When excluding the first case as an outlier, the correlation was slightly lower but still strong at 0.68 ($p < 0.001$).

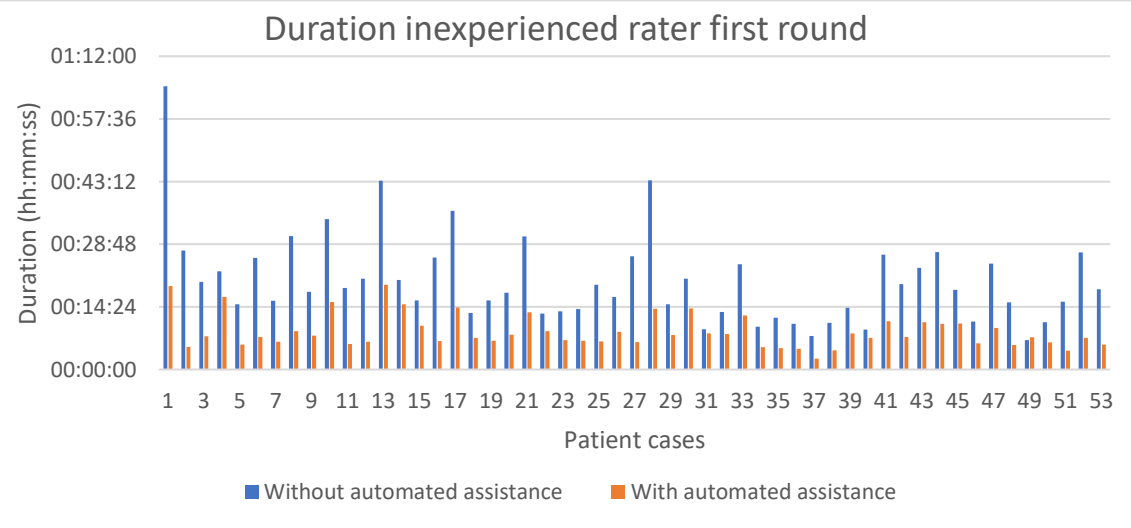

Figure 25: Time spent evaluating each case in the first round of 53 cases by the inexperienced rater, with AI assistance shown in orange and without AI assistance shown in blue. Each bar represents the time taken to evaluate a single case.

A paired t-test was conducted to statistically analyze the evaluation duration of the second round with the remaining 57 patient cases. The test revealed a statistically significant difference in duration of 02:46 minutes (95% CI: 01:17 min – 04:13 min) when comparing the evaluation without automated assistance (M = 09:59 min, SD = 05:52 min) to the assessment with automated support (M = 07:13 min, SD = 02:57 min), $t(56) = 3.75$, $p < 0.001$. The effect size, calculated using Cohen's d, was lower than in the first evaluation round, with $d = 0.50$ (95% CI: 0.22 – 0.77), indicating a medium effect size.

For the second round with 57 patient cases, a Pearson correlation was calculated to assess the relationship between evaluation duration without and with use of the automated program, resulting in a moderate correlation of 0.36 ($p < 0.006$). Thus, the correlation in the second round was weaker than in the first round.

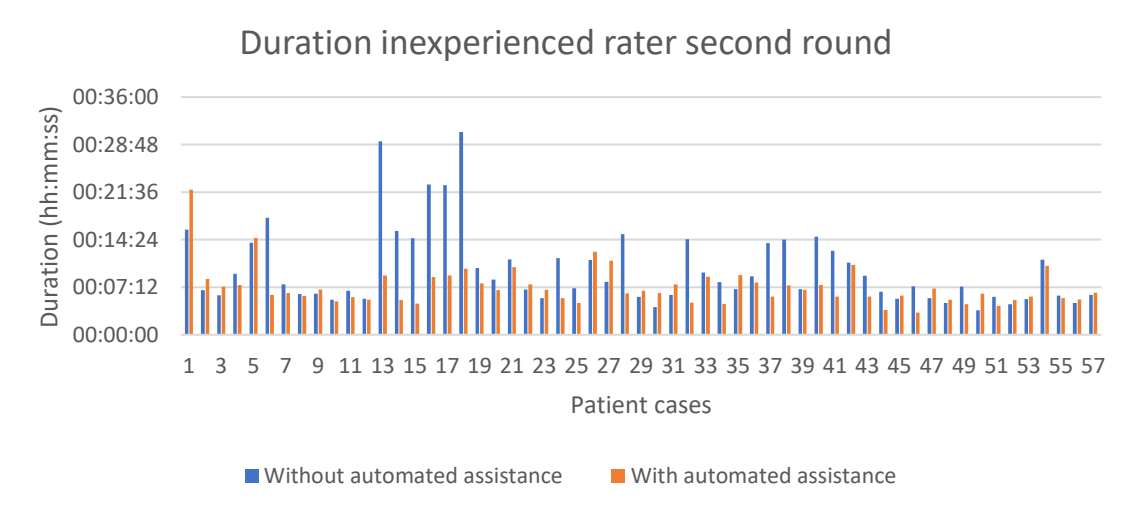

Figure 26: Duration for evaluating each of the 57 cases in the second round by the inexperienced rater, displayed in orange with AI assistance and in blue without AI assistance. Each bar shows the evaluation time for an individual case.

4.5 Summary of the Results

At the case level, neither the experienced nor the inexperienced rater showed statistically significant differences in sensitivity, specificity, PPV, NPV or overall accuracy when comparing evaluations with and without the automated program. Notably, the inexperienced rater consistently demonstrated higher sensitivity and NPV than the experienced rater, both with and without the AI-based software.

The interrater reliability measured by unweighted Cohen's κ improved with automated assistance, rising from κ = 0.56 without automated support to κ = 0.68 with it.

At the lesion level, both raters showed improved sensibility, PPV and F1 scores with the automated tool. Interrater reliability, again using unweighted Cohen's κ, slightly increased from κ = 0.61 without the software to κ = 0.64 with its use.

Additionally, the volume of all progressive lesions was analyzed, with separate assessments for new and enlarged lesions. Across all raters and software conditions, detected lesions were significantly larger than undetected ones. For newly detected lesions, while the inexperienced rater identified smaller lesions on average compared to the experienced rater, both with and without automated assistance, this volume difference was not statistically significant. Similarly, no statistically significant volume differences were found for undetected lesions between the two raters, regardless of software assistance.

The assessment of enlarged lesions by both raters as well as the AI-based software showed that lesions classified as progressive had larger median volume increases than the undetected lesions, though none of the approaches showed a statistically significant difference in these volume differences. Both with and without software assistance, the inexperienced rater identified enlarged lesions at smaller volume increases than the experienced rater, though these differences were not statistically significant as well.

Moreover, evaluation duration was measured. Since only the absolute time was recorded for the experienced rater, no statistical analysis could be performed. The average time per case was reported as 05:15 minutes without AI-based assistance and 03:14 minutes with assistance, which refers to a percentual reduction of 38.41%.

For the inexperienced rater, a statistically significant reduction in evaluation time per case was found, decreasing from 14:31 minutes without automated assistance to 07:50 minutes with it, which was a percentual reduction of 46.04%. The first case was excluded from the analysis due to a potential learning effect.

Additionally, a significant reduction in the average duration was observed across both rounds when using the evaluation program. However, the effect in the second round was not as strong as in the first round.

5 Discussion

This study aimed to assess the evaluation of MRI images of MS patients by raters with different levels of experience, both with and without the support of an automated lesion detection tool. Tracking lesion activity to monitor MS disease progression has emerged as an important factor in determining whether adjustments to immunomodulatory therapy are necessary (Weiner 2009). However, AI-assisted methods for assessing MS lesion activity using baseline and follow-up MRI scans remain limited, and only a few studies have investigated how radiologists and medical staff with no prior radiological experience perform with the support of automated lesion detection systems compared to without such assistance.

Therefore, this study focuses on the evaluation of longitudinal MRI data from MS patients, with raters being supported by an automated lesion detection program that highlights unchanged, new, enlarged, or reduced lesions in distinct colors.

The primary objective was to determine how the use of this tool affects the classification of follow-up MRI images in terms of progressive versus non-progressive disease courses, based on the identification of new or enlarged lesions, across raters with varying levels of expertise. Additionally, the study aimed to investigate whether the use of this automated tool provides advantages in terms of evaluation duration and interrater variability.

5.1 Impact of automated assistance on the assessment of disease course and lesion detection

The results of this study indicate that the use of an automated lesion detection system did not yield a statistically significant improvement in sensitivity,

specificity, PPV, NPV, or overall accuracy in assessing whether a patient's disease course was progressive or stable, regardless of the rater's level of radiological experience.

At the lesion level, both experienced and inexperienced raters benefited from the use of the automated tool, showing improvements in sensitivity, PPV, and F1 scores. Nevertheless, these improvements did not reach statistical significance in the lesion-level analysis. However, this enhancement in lesion detection did not translate into an improvement in correctly identifying progressive cases as progressive, which implies that detecting more lesions does not necessarily lead to a more accurate assessment of disease course.

Interestingly, the inexperienced rater consistently demonstrated significantly higher sensitivity, NPV and accuracy in assessing the longitudinal MRI data as either progressive or non-progressive compared to the experienced rater both with and without automated assistance. One possible explanation for the higher sensitivity observed in the inexperienced rater could be the extended evaluation time, allowing the identification of a greater number of progressive lesions. Another hypothesis is that, in clinical practice, the primary focus is on determining whether disease progression is present rather than precisely quantifying and localizing each progressive lesion. Consequently, experienced raters may adopt a more pragmatic approach, potentially ceasing detailed evaluation after identifying a sufficient number of lesions indicative of disease progression.

Contrary to our findings, previous studies have reported improvements in lesion detection and disease course classification when employing AI-assisted workflows. For instance, Combès et al. (Combès et al. 2021) demonstrated that integrating an automated lesion detection workflow significantly increased the number of identified new MS lesions in longitudinal MRI scans, even among highly experienced neuroradiologists, and influenced the classification of patient's disease course as stable or active based on MRI findings. In their study, the sensitivity of three neuroradiologists for detecting new lesions improved substantially with AI assistance, increasing from 0.60 – 0.66 without support to 0.75 – 0.84 with AI, and all improvements were statistically significant. Similarly,

Van Hecke et al. (van Hecke et al. 2021) found that incorporating AI-assisted radiological reading by the tool icobrain MS significantly increased the proportion of patients classified as having disease activity compared to visual assessment alone. Their findings suggest that AI augmentation can enhance radiological detection of disease activity, potentially impacting clinical decision-making.

Considering these contrasting results, the discrepancies between our findings and previous studies may be attributable to differences in study design, the specific AI-assisted tool employed, or the complexity of lesion interpretation in real-world clinical settings. Additionally, the baseline sensitivity for detecting disease progression without automated assistance was already relatively high compared to the other studies, at 0.7 for the experienced rater and 0.95 for the inexperienced rater, leaving limited room for further improvement using AI-based software in this context.

While prior research highlights the potential for automated lesion detection to enhance lesion identification rates, our study suggests that this benefit does not necessarily translate into improved overall diagnostic performance regarding disease progression assessment.

Further analysis of lesion volume dynamics revealed that the detected progressive lesions were significantly larger than undetected ones, regardless of the rater or software condition. Similarly, no significant differences in the volume of undetected lesions were found between the raters, suggesting that lesion detection was primarily size-dependent rather than influenced by experience or software use. This implies that while AI-based tools may support lesion detection, their influence on volumetric assessment remains limited.

Our findings align with prior studies which highlight the influence of lesion volume on detection rates. Commowick et al. (Commowick et al. 2018) emphasize that lesion detection rates significantly decline as lesion volume decreases, leading to false negatives in automated segmentation. The link between lesion volume and detection accuracy is further supported by Cabezas et al. (Cabezas et al. 2016), who state that small lesions between 3 and 10 voxels had a notably low

detection rate of only 42.86% compared to a detection rate of 77.42% for large lesions of >50 voxels.

Overall, these findings highlight the limitations of current MS lesion detection methods, both manual and automated, in identifying small lesions. This poses a challenge in monitoring MS progression, as newly formed lesions, which are often small, may be overlooked (Basaran et al. 2022) and emphasized the need for advancements in imaging techniques and algorithmic sensitivity to improve early lesion detection.

5.2 Use of automated assistance for evaluators with varying radiological experience

Another key aspect of this study was to assess how raters with different levels of radiological experience benefit from an AI-based lesion detection tool.

In this study, all raters, regardless of their expertise in neuroradiological imaging, exhibited higher sensitivity, PPV, and F1 scores at the lesion level, though the improvements were not statistically significant. However, the improved lesion detection did not translate into significantly better identification of disease course. Notably, the inexperienced rater consistently demonstrated higher sensitivity, NPV and accuracy than the experienced raters, regardless of automated assistance. Therefore, the findings of our study support the implementation of the AI-based tool from the beginning of specialist medical training even among junior physicians without prior neuroradiological experience, as its use required no specific training.

Previous research has investigated the influence of radiological expertise on MS lesion detection, with some studies specifically examining the role of AI support in this context. Contrary to our findings, Wang et al. (Wang et al. 2017) demonstrated that neuroradiologists outperformed non-neuroradiologists in conventional MS imaging analysis, achieving a sensitivity of 82% compared to 42% in the non-neuroradiologist group, as well as a higher NPV of 89% vs. 64%, and a lower FNR of 18% vs. 58%. Similarly, Combès et al. (Combès et al. 2021)

observed that experienced neuroradiologists generally detected more new lesions than less experienced clinicians, although these differences were not statistically significant. However, their study also found that the implementation of AI-generated lesion masks led to a more than 15% increase in lesion detection across all raters, indicating that AI provides benefits irrespective of experience level. Likewise, Dahan et al. (Dahan et al. 2018) reported that both trained and untrained raters exhibited a statistically significant improvement in new lesion detection when using AI assistance. While this aligns with our findings at the lesion level, it diverges from our case-level results, where overall diagnostic accuracy remained unaffected. These varying study results emphasize the need for additional research to determine whether AI tools can compensate for limited experience in lesion detection.

5.3 Impact of AI-based evaluation programs on interrater reliability

An additional objective of our study was to assess interrater reliability and evaluate the impact of AI-based assistance on agreement between evaluators. The use of automated assistance improved interrater reliability in disease course assessment, as indicated by an increase in Cohen's κ from 0.56 (moderate agreement) to 0.68 (substantial agreement). This suggests that AI-based software enhances consistency between raters, even in the absence of significant changes in overall diagnostic performance. At the lesion level, interrater reliability also slightly increased from 0.61 to 0.64 with software use, indicating more consistent lesion classification. These findings support the potential of AI tools to reduce interrater variability.

Our findings align with previous studies that have investigated interrater reliability in similar contexts. Park et al. (Park et al. 2022) reported moderate interrater agreement (Cohen's κ = 0.567) for standard image review, which improved to almost perfect agreement (Cohen's κ = 0.822) with the use of color complement image maps. Another study by Combès et al. (Combès et al. 2021) also reported a reduction in inter-expert variability when segmentation masks were used, with the overall Fleiss's Kappa coefficient decreasing from 0.59 (95% CI: 0.49 – 0.69)

without AI support to 0.47 (95% CI: 0.38 – 0.57) with AI. However, this reduction did not reach statistical significance. Van Hecke et al. (van Hecke et al. 2021) demonstrated a 32.5% improvement in interrater lesion count agreement with AI-assisted annotations, reinforcing the notion that automated support can enhance consistency in lesion detection. In contrast, Peters et al. (Peters et al. 2024) found that the interclass correlation coefficient was already high without automated assistance (0.827 in cohort 1 and 0.922 in cohort 2) and did not change significantly with AI support (0.837 in cohort 1 and 0.820 in cohort 2), suggesting that in settings with already high reliability, AI assistance may not provide additional benefits.

In conclusion, while AI-based assistance can improve interrater reliability in disease course assessment and lesion classification, its impact might be more pronounced in scenarios where baseline agreement is moderate rather than already high. Despite these improvements, visual MRI reading remains prone to interrater variability, especially in high lesion burden cases (Galletto Pregliasco et al. 2018). In our study, raters focused on accuracy rather than routine clinical MRI reading. Therefore, in clinical practice, time constraints may further increase variability, emphasizing the need for further research to evaluate the performance of AI tools in real-world clinical settings.

5.4 Impact of AI-based lesion detection programs on evaluation time

Another key aspect of the study was the impact of automated assistance on evaluation duration. Interpreting brain MRIs of MS patients can be time-consuming, especially in cases with multiple lesions (Galletto Pregliasco et al. 2018), as accurate tracking of lesion changes is crucial. In our study, due to the absence of individual time recordings for the experienced raters, statistical analysis could not be performed. However, the experienced rater's reported times suggest a reduction from 05:15 minutes without AI to 03:14 minutes with AI per case, a 38.41% decrease. For the inexperienced rater, a significant reduction was observed, from 14:31 minutes to 07:50 minutes per case, which reflects a reduction in reading time by 46.04%.

The exclusion of the first case for the inexperienced rater, due to a potential learning effect, strengthens the argument that the observed efficiency gain was attributed to AI assistance rather than improvements in performance through practice alone. To further detect any learning effect for the inexperienced rater, the 110 cases were evaluated in two consecutive rounds. It was found that, even after excluding the first case as an outlier in the first round of 53 cases, the inexperienced rater was still significantly slower in the first round compared to the second round of 57 cases. Notably, when supported by the AI tool, the inexperienced rater was faster in the second round than in the first as well, though this difference did not reach statistical significance. Additionally, the effect size of the AI tool in the second round was smaller than in the first round, which may suggest the presence of a learning effect. This is particularly relevant, as the evaluation in each round with AI support was conducted after the evaluation without AI.

Ultimately, the AI-based tool demonstrated a beneficial effect on both experienced and inexperienced raters in terms of reducing evaluation time. However, this reduction in time may not be as pronounced for raters without prior radiological experience as the overall improvement seen in the inexperienced rater might suggest. The real-world effect is likely closer to the size observed in the second round, indicating that while the tool is effective for both groups, its impact may diminish as raters gain more experience with its use.

Multiple previous studies align with our findings. Galletto Pregliasco et al. (Galletto Pregliasco et al. 2018) reported a significant reduction in reading times, decreasing from 192 seconds per case with conventional evaluation to 106 seconds, including processing time, when using automated MRI co-registration. Similarly, Van Hecke et al. (van Hecke et al. 2021) observed that AI-assisted radiological reporting shortened evaluation duration from 07:28 minutes to 05:49 minutes, improving efficiency and enabling approximately 13 reports per hour compared to 8 with traditional methods, a 40% increase. Rathmann et al. (Rathmann et al. 2024) showed a reduction in assessment time by 210 seconds with machine learning, along with a more consistent distribution of examination times. Peters et al. (Peters et al. 2024) found that AI-supported image analysis

significantly reduced reporting time, with benefits depending on radiologist experience. The reporting time for the lesion load assessment dropped from 286.85 to 196.34 seconds ($p > 0.001$), and for follow-up scans from 196.17 to 120.87 seconds ($p < 0.001$). Sieber et al. (Sieber et al. 2024) reported a statistically significant average reduction of 2.83 minutes in the reading times of resident radiologists per MS-MRI evaluation. Likewise, Combès et al. (Combès et al. 2021) observed a substantial decrease of 02:45 minutes in reporting times with the use of AI tools.

One possible explanation for the significantly shorter evaluation times is that AI-assisted tools highlight lesion changes, allowing radiologists to filter out unchanged lesions and instead focus on newly developed or enlarged lesions (Cabezas et al. 2016). Incorporating ML software into the MS MRI workflow may enhance decision-making by guiding attention to critical areas, thereby making manual assessment more efficient (Rathmann et al. 2024).

Our study offers valuable evidence that AI-based decision support can save time for neuroradiologists and healthcare personnel without prior radiological experience. As radiological workloads increase, these time savings can be allocated to tasks where AI assistance is not yet available (Sieber et al. 2024; Waller et al. 2022).

5.5 Implementation of AI-based assistance in clinical routine and associated costs

The integration of AI-based tools into clinical routine remains a key challenge in neuroradiological imaging, particularly regarding their practical applicability in daily workflows. Our study demonstrated that an AI-based program can improve the efficiency of evaluating MRI data in MS patients by accelerating the assessment of disease progression. However, it has not yet been tested in routine clinical practice. Therefore, further research is essential to determine how well this tool integrates into the clinical workflow and whether it offers a meaningful advantage in routine practice.

One of the main challenges in clinical routine is the increasing workload combined with limited personnel resources (Rathmann et al. 2024; Peters et al. 2024). Our study provides strong evidence that AI-based tools can effectively address this issue by significantly accelerating the evaluation of MRI data in MS patients and reinforces the potential of AI to enhance diagnostic efficiency while maintaining high-quality assessments. Therefore, the use of AI could enable radiologists to dedicate more time to essential tasks such as interpreting results, coordinating examinations, or engaging in research (Sieber et al. 2024; Waller et al. 2022). Furthermore, our results highlight that AI solutions can increase patient throughput by allowing more examinations to be conducted within the same timeframe. This not only supports overburdened radiology departments but also strengthens the economic justification for AI integration in clinical practice (Rathmann et al. 2024), making it a valuable tool for optimizing healthcare resource management.

Although AI-based tools offer promising benefits, a considerable gap remains between their development and actual implementation in clinical practice, as few studies have integrated them into clinical workflows, and even fewer have evaluated their application in clinical settings (Spagnolo et al. 2023). To date, there is no consistent evidence which supports the successful and widespread adoption of AI in clinical neuroradiology (Barnett et al. 2023; Bonacchi et al. 2022).

However, some studies demonstrate promising results. Van Hecke et al. (van Hecke et al. 2021) tested the icobrain ms software, an AI-based MRI analysis tool, in clinical practice and found that it improved radiological workflow efficiency by 40%, significantly reducing intra- and interrater variability. Similarly, Dahan et al. (Dahan et al. 2019) reported that implementing semi-automated imaging software in their department was largely seamless, with no significant increase in picture archiving and communication system transfer times or data storage requirements. Moreover, staff surveys indicated overwhelmingly positive feedback regarding its integration into daily practice.

Admittedly, the successful implementation of AI into clinical routine requires extensive training, validation, and close cooperation between medical professionals and AI specialists (Naji et al. 2023), a process that can be time-consuming, complex and costly. Therefore, the implementation of AI in MS lesion segmentation presents several financial challenges. Despite the potential benefits, there is currently a lack of studies specifically addressing the expenses involved in developing and implementing AI-based segmentation tools in clinical practice (Pham et al. 2023; Winder et al. 2024), as most AI research on medical imaging focuses on algorithm performance rather than healthcare costs and patient outcomes.

A microsimulation study by Sima et al. (Sima et al. 2021) evaluated the health-economic impact of AI-assisted treatment decisions in MS. Their findings indicate that incorporating AI-supported MRI lesion assessment into clinical decision-making could generate estimated savings of $2,155 per patient annually after 10 years and $2,267 after 15 years when MRI lesion and brain atrophy criteria are added to conventional clinical assessments. However, these potential savings must be weighted against MRI scan and radiological evaluation costs, which range from $2,000 to $4,000 per year.

Although not directly related to MS imaging, an early health technology assessment by Kicky G. van Leeuwen et al. (Kicky G. van Leeuwen et al.) on AI-aided vessel occlusion detection in acute stroke provides insight into AI's economic impact in neuroradiology. Their findings showed that AI-assisted diagnosis reduced missed cases and therefore has the potential to improve healthcare outcomes and lead to long-term cost savings. The cost-effectiveness of AI was evaluated by comparing the per-use cost of AI analysis with the projected reductions in future medical expenditures resulting from improved clinical outcomes. However, a comprehensive financial assessment should also consider initial investments in software and infrastructure, ongoing operational costs and training expenses (Wolff et al. 2020).

Despite increasing research on AI in radiology, a key challenge remains the lack of high-quality literature confirming that AI-based segmentation tools justify the

financial risk (Winder et al. 2024). Without clear evidence, healthcare providers may hesitate to invest in AI integration despite its potential benefits.

Overall, while AI-based tools show potential for improving efficiency and diagnostic consistency, their full integration into clinical routine remains an ongoing challenge. The limited number of real-world studies highlights the need for further research to validate their reliability and effectiveness. For AI-based lesion segmentation tools to be effectively applied, they must seamlessly integrate with existing clinical workflows and handle MRI data from different sources without preprocessing. Future research should prioritize the development of AI solutions that seamlessly integrate into existing workflows while ensuring user acceptance, clinical utility, and cost-effectiveness.

5.6 Limitations

This study has several limitations that should be considered when interpreting the results.

First, the study was conducted under controlled conditions without the time and cost pressures typically present in clinical practice. The absence of these external pressures in our study setting may have led to a more thorough assessment of the MRI images for new or enlarged lesions, potentially resulting in higher sensitivity and better interrater reliability than what might be observed in routine clinical practice. This consideration is particularly relevant for the inexperienced rater, as the experienced raters performed the assessment in addition to their regular clinical duties. This circumstance may have also contributed to their relatively low performance compared to other studies.

Additionally, the MRI data was acquired from different scanners, which reflects clinical reality. However, only two different scanner types with 3T field strengths were used, making it unclear whether the findings can be generalized to a broader range of MRI systems with varying field strengths and acquisition protocols. This is particularly relevant given that higher magnetic field strengths have been shown to improve the detection of progressive MS lesions (Filippi et

al. 1995). Consequently, the use of different MRI scanners may have influenced the study results in terms of sensitivity and interrater variability.

Furthermore, the side-by-side comparison of MRI data for detecting progressive MS lesions was performed only on FLAIR and T1w images, as these sequences are commonly used in clinical practice for lesion identification and monitoring (Gramsch et al. 2015; Roy et al. 2018). However, other MRI sequences, such as T2w or PDw scans, may provide additional diagnostic information and influence lesion detectability (Simon et al. 2006). As a result, our findings may not be applicable to differently weighted images, which could potentially impact sensitivity and interrater variability.

Another potential limitation is that not all raters evaluated the MRI images on the same monitor, which means that variations in display quality could have influenced lesion detection and segmentation performance. Since subtle differences in image presentation can impact the visibility of lesions, particularly smaller or less well-defined ones, this variability may have contributed to differences in sensitivity and interrater reliability. However, this also reflects real-world clinical practice, when radiologists often work with different display systems.

Besides that, the volumetric assessment performed in 3D Slicer, the software used for measuring lesion volumes, is not completely objective and is further influenced by the slice thickness of the MRI scans (Filippi et al. 1995). These factors may introduce variability among raters, particularly when segmenting new or enlarged MS lesions. Nonetheless, this variability is also representative of real-world clinical practice, where differences between raters are frequently observed, as lesion segmentation remains a subjective process (Altay et al. 2013; Egger et al. 2017).

Another limitation of this study is the variation in the cut-off values used to define a progressive lesion across different methods. For the AI tool, a lesion was considered progressive if there was a 30% increase in volume compared to the previous MRI scan, while in the side-by-side comparison, a 50% increase in volume was required for a lesion to be classified as progressive. Additionally, the

ground truth was established through a purely visual comparison by the neuroradiologist with 16 years of experience. As a result, a lesion with a 40% volume increase according to the volumetric assessment in 3D Slicer was considered progressive, potentially leading to discrepancies between the methods used to classify lesion progression. These varying thresholds could impact the consistency and reliability of the results when comparing the effectiveness of the manual assessment with the evaluation supported by the AI tool.

Moreover, blinding to the use of AI support was not possible, as raters were aware of whether they were utilizing AI assistance. This introduces the potential concern that AI-supported evaluations may have been performed more quickly simply due to the knowledge of AI involvement. However, despite the faster evaluation times, there was no significant reduction in the quality of progressive lesion detection. This suggests that the reduced evaluation time was not solely a result of the perceived efficiency gain associated with the use of AI support. Nonetheless, this effect cannot be ruled out definitively.

For the inexperienced rater, the observed reduction in evaluation time with AI support could also, in part, be attributed to a learning effect rather than AI assistance alone. While the first case was excluded from the evaluation of assessment time to minimize a potential learning effect, it cannot be excluded that learning still influenced the evaluation duration. Additionally, the splitting of the evaluations in two rounds revealed that, despite the exclusion of the first case, the inexperienced rater remained slower in the first round compared to the second, although the difference was not statistically significant with AI support. Thus, while use of the AI tool improved evaluation duration, the reduction in time may not be as pronounced for raters without prior radiological experience as suggested by the overall reduction in evaluation duration for the inexperienced rater but is likely closer to the effect observed in the second round. Future studies could address this limitation by implementing an alternative evaluation design, in which AI-assisted analysis is performed both prior to and following conventional side-by-side comparison.

Lastly, the rater cohort in this study was relatively small, consisting of two experienced and one inexperienced rater, which may not fully represent the diversity of expertise typically found in clinical practice. Additionally, the study population included only 110 cases, which may not be large enough to account for the full range of variability seen in real-world clinical settings.

Despite these limitations, this study provides valuable insights into the impact of AI-assisted lesion segmentation among raters with different levels of neuroradiological experience. Future studies with a more diverse range of scanner types, a larger rater cohort with varying expertise, and a larger patient population under real clinical conditions will be necessary to further validate our findings.

5.7 Conclusion

In conclusion, the results of our study suggest that AI-assisted MRI lesion detection offers significant potential to enhance the efficiency of assessing disease activity in MS patients. The integration of AI can reduce evaluation time without compromising the detection of progressive cases, regardless of the evaluator's radiological experience. This is especially important given the growing workload in radiology departments, where AI-based automation could contribute to optimizing resource allocation (Waller et al. 2022; Sieber et al. 2024). Our results underscore the considerable potential of integrating AI-based tools into clinical practice for the assessment of disease activity in MS patients, highlighting their capacity to provide both rapid and accurate assessments, even for junior physicians without prior neuroradiological experience.

6 Summary

MS is a chronic neurodegenerative disease which causes focal lesions in the CNS, leading to gradual disability (Steinman 1996; Compston and Coles 2008). Cranial MRI plays a crucial role in its diagnosis and monitoring of disease activity by assessing lesion load and identifying new or enlarged lesions (Polman et al. 2011; Filippi et al. 2018). In clinical practice, lesion progression is typically evaluated through side-by-side comparison of sequential MRI scans, a method that is both time-consuming and prone to errors (Altay et al. 2013; Egger et al. 2017). Therefore, automated detection systems to enhance efficiency and diagnostic accuracy are highly desirable.

This study investigated the impact of AI-assisted MS lesion detection on the evaluation of MRI scans by raters with different levels of radiological experience. The study was conducted by three raters, two experienced neuroradiologists (16 and 8 years of experience) and one inexperienced medical student. Each rater assessed the data manually and with AI software (AIRAscore) to identify new or enlarged lesions. The results indicated that AI-support improved sensitivity, PPV, and F1 scores at the lesion level for both experienced and inexperienced raters. However, no statistically significant improvement was observed in the classification of disease progression at the case level. Interestingly, the inexperienced rater consistently demonstrated higher sensitivity and NPV than the experienced rater, both with and without AI assistance.

One of the key advantages of AI-based assessment was the reduction in evaluation time. The use of the automated tool resulted in a time reduction from 05:15 minutes to 03:14 minutes per case for the experienced raters and a statistically significant decrease for the inexperienced rater, from 14:31 minutes per case to 07:50 minutes. Additionally, interrater reliability improved with automated assistance, suggesting that AI tools can enhance consistency in lesion detection and classification, even when no substantial gains in diagnostic accuracy are observed.

Looking ahead, AI-assisted MRI lesion detection holds great potential for enhancing the efficiency of the assessments by reducing the evaluation time

without compromising the detection rate of progressive cases, regardless of the evaluator's prior radiological experience. This efficiency gain is particularly relevant given the increasing workload in radiology departments (Peters et al. 2024), where AI-driven automation could help optimize resource allocation and allow neuroradiologists to focus on more complex diagnostic and therapeutic decisions (Sieber et al. 2024; Waller et al. 2022).

However, despite the promising benefits, the widespread clinical adoption of AI-assisted MS lesion detection faces several challenges. The costs associated with AI implementation, including software acquisition, infrastructure requirements and training (Wolff et al. 2020), must be carefully weighted against the time savings and potential improvements in diagnostic consistency and the option for quicker therapeutic adjustments resulting from earlier detection of progressive disease courses (Weiner 2009).

Future research should focus on validating the clinical utility of AI tools in real-world settings, ensuring seamless integration into existing workflows, and assessing their long-term cost-effectiveness. As AI technology continues to evolve, its role in neuroradiology is expected to grow, with the potential to improve patient care by enabling faster, more standardized, and more reliable disease monitoring (Bonacchi et al. 2022). The results of our study support the potential benefits of implementing AI-based tools for the assessment of disease activity in patients with MS, demonstrating efficacy not only for experienced neuroradiologists but also for junior physicians without prior neuroradiological experience early in their specialist training.

7 Deutsche Zusammenfassung

Multiple Sklerose ist eine chronische neurodegenerative Erkrankung, die fokale Läsionen im zentralen Nervensystem verursacht und zu schrittweiser Behinderung führen kann (Steinman 1996; Compston und Coles 2008). Die cerebrale MRT-Bildgebung spielt eine entscheidende Rolle bei der Diagnose und der Überwachung der Krankheitsaktivität, indem anhand dieser die Läsionslast bewertet wird und neue oder vergrößerte Läsionen identifiziert werden können (Polman et al. 2011; Filippi et al. 2018). In der klinischen Praxis wird das Auftreten progredienter Läsionen typischerweise durch den Vergleich von zeitlich aufeinanderfolgenden MRT-Bildern bewertet, was sowohl zeitaufwändig als auch fehleranfällig ist (Altay et al. 2013; Egger et al. 2017). Daher sind automatisierte Erkennungssysteme zur Steigerung der Effizienz und diagnostischen Genauigkeit wünschenswert.

Diese Studie untersuchte den Einfluss der durch künstliche Intelligenz (KI) unterstützten MS-Läsionserkennung auf die Bewertung einer möglichen Krankheitsprogredienz anhand von MRT-Scans durch Auswertende mit unterschiedlicher radiologischer Vorerfahrung. Die Studie wurde von drei Ratern durchgeführt: zwei Neuroradiologen (16 und 8 Jahre Erfahrung) und einer Medizinstudentin ohne radiologische Vorerfahrung. Jeder Beurteilende bewertete die Daten manuell und mit Unterstützung durch eine KI-Software (AIRAscore), um neue oder vergrößerte Läsionen zu identifizieren. Die Ergebnisse zeigten, dass mit KI-Unterstützung die Sensitivität, der positive prädiktive Wert und der F1-Wert auf Läsionsebene sowohl für die erfahrenen als auch für den unerfahrenen Rater verbessert wurden. Es wurde jedoch keine statistisch signifikante Verbesserung für die Klassifikation des Erkrankungsfortschritts auf der Fall-Ebene beobachtet. Interessanterweise zeigte die Medizinstudentin ohne radiologische Vorerfahrung sowohl mit als auch ohne KI-Unterstützung eine konsequent höhere Sensitivität und einen höheren negativen prädiktiven Wert als die erfahrenen Rater.

Ein wesentlicher Vorteil der KI-gestützten Auswertung war die Reduktion der Bewertungszeit. Der Einsatz des automatisierten Programms führte zu einer

Zeitreduktion von 05:15 Minuten auf 03:14 Minuten pro Fall für die erfahrenen Rater und einer statistisch signifikanten Verringerung von 14:31 Minuten pro Fall auf 07:50 Minuten für den unerfahrenen Rater. Darüber hinaus verbesserte sich die Interrater-Reliabilität mit automatisierter Unterstützung, was darauf hindeutet, dass der Einsatz von KI die Läsionserkennung und -klassifikation verbessern kann, selbst wenn keine wesentliche Verbesserung in der diagnostischen Genauigkeit auf Fallebene beobachtet werden konnte.

Blickt man in die Zukunft, so bietet die KI-unterstützte MRT-Läsionserkennung großes Potenzial zur Effizienzsteigerung, indem mit KI die Auswertungszeit verkürzt werden kann, ohne die Erkennungsrater progressiver Fälle zu beeinträchtigen – unabhängig von der radiologischen Erfahrung des Auswertenden. Dieser Effizienzgewinn ist besonders relevant angesichts der zunehmenden Arbeitsbelastung in radiologischen Abteilungen (Peters et al. 2024), in denen der Einsatz von KI dazu beitragen könnte, die Ressourcenverteilung zu optimieren und er Neuroradiologen zu ermöglichen, sich auf komplexere diagnostische und therapeutische Entscheidungen zu konzentrieren (Sieber et al. 2024; Waller et al. 2022).

Trotz der vielversprechenden Vorteile stehen der breiten klinischen Einführung der KI-unterstützen MS-Läsionserkennung mehrere Herausforderungen gegenüber. Die mit der Implementierung von KI verbundenen Kosten, einschließlich Softwareerwerb, Infrastrukturanforderungen und Schulungen (Wolff et al. 2020), müssen sorgfältig gegen die Zeitersparnisse und potenziellen Verbesserungen der diagnostischen Beständigkeit sowie die Möglichkeit einer schnelleren therapeutischen Anpassung durch frühere Erkennung progressiver Krankheitsverläufe (Weiner 2009) abgewogen werden.

Zukünftige Forschung sollte sich darauf konzentrieren, den Nutzen von KI-Instrumenten im realen klinischen Setting zu validieren, ihre nahtlose Integration in bestehende Arbeitsabläufe sicherzustellen und ihre langfristige Kosteneffizienz zu bewerten. Da sich die KI-Technologie weiterhin entwickelt, wird ihre Rolle in der Neuroradiologie voraussichtlich wachsen. Hieraus ergibt sich das Potenzial, die Patientenversorgung zu verbessern, indem der Einsatz von KI eine

schnellere, standardisierte und zuverlässigere Krankheitsüberwachung ermöglicht (Bonacchi et al. 2022). Die Ergebnisse unserer Studie unterstützen die potenziellen Vorteile des Einsatzes von KI-basierten Instrumenten zur Beurteilung der Krankheitsaktivität bei Patienten mit MS und zeige dies sowohl für erfahrene Neuroradiologen als auch für junge Ärzte ohne vorherige neuroradiologische Erfahrung zu Beginn ihrer Facharztausbildung.

8 Literaturverzeichnis

Abraham, Alison G.; Duncan, Donald D.; Gange, Stephen J.; West, Sheila (2009): Computer-aided assessment of diagnostic images for epidemiological research. In: *BMC medical research methodology* 9, S. 74. DOI: 10.1186/1471-2288-9-74.

Afzal, H. M. Rehan; Luo, Suhuai; Ramadan, Saadallah; Lechner-Scott, Jeannette (2022): The emerging role of artificial intelligence in multiple sclerosis imaging. In: *Multiple sclerosis (Houndmills, Basingstoke, England)* 28 (6), S. 849–858. DOI: 10.1177/1352458520966298.

Altay, Edru Erbeyat; Fisher, Elizabeth; Jones, Stephen E.; Hara-Cleaver, Claire; Lee, Jar-Chi; Rudick, Richard A. (2013): Reliability of classifying multiple sclerosis disease activity using magnetic resonance imaging in a multiple sclerosis clinic. In: *JAMA neurology* 70 (3), S. 338–344. DOI: 10.1001/2013.jamaneurol.211.

Andravizou, Athina; Dardiotis, Efthimios; Artemiadis, Artemios; Sokratous, Maria; Siokas, Vasileios; Tsouris, Zisis et al. (2019): Brain atrophy in multiple sclerosis: mechanisms, clinical relevance and treatment options. In: *Auto-immunity highlights* 10 (1), S. 7. DOI: 10.1186/s13317-019-0117-5.

Ascherio, Alberto (2013): Environmental factors in multiple sclerosis. In: *Expert review of neurotherapeutics* 13 (12 Suppl), S. 3–9. DOI: 10.1586/14737175.2013.865866.

Aslani, Shahab (2020): Deep Learning Approaches for The Segmentation of MS lesions on Brain MRI.

Aslani, Shahab; Dayan, Michael; Storelli, Loredana; Filippi, Massimo; Murino, Vittorio; Rocca, Maria A.; Sona, Diego (2019): Multi-branch Convolutional Neural Network for Multiple Sclerosis Lesion Segmentation. In: *NeuroImage* 196 (5), S. 1–15. DOI: 10.1016/j.neuroimage.2019.03.068.

Bakshi, R.; Ariyaratana, S.; Benedict, R. H.; Jacobs, L. (2001): Fluid-attenuated inversion recovery magnetic resonance imaging detects cortical and

juxtacortical multiple sclerosis lesions. In: *Archives of neurology* 58 (5), S. 742–748. DOI: 10.1001/archneur.58.5.742.

Barile, Sergio; Piciocchi, Paolo; Bassano, Clara; Spohrer, Jim; Pietronudo, Maria Cristina (2019): Re-defining the Role of Artificial Intelligence (AI) in Wiser Service Systems. In:. International Conference on Applied Human Factors and Ergonomics: Springer, Cham, S. 159–170. Online verfügbar unter https://link.springer.com/chapter/10.1007/978-3-319-94229-2_16.

Barkhof, Frederik; Simon, Jack H.; Fazekas, Franz; Rovaris, Marco; Kappos, Ludwig; Stefano, Nicola de et al. (2011): MRI monitoring of immunomodulation in relapse-onset multiple sclerosis trials. In: *Nature reviews. Neurology* 8 (1), S. 13–21. DOI: 10.1038/nrneurol.2011.190.

Barnett, Michael; Wang, Dongang; Beadnall, Heidi; Bischof, Antje; Brunacci, David; Butzkueven, Helmut et al. (2023): A real-world clinical validation for AI-based MRI monitoring in multiple sclerosis. In: *NPJ digital medicine* 6 (1), S. 196. DOI: 10.1038/s41746-023-00940-6.

Bar-Or, Amit; Grove, Richard A.; Austin, Daren J.; Tolson, Jerry M.; VanMeter, Susan A.; Lewis, Eric W. et al. (2018): Subcutaneous ofatumumab in patients with relapsing-remitting multiple sclerosis: The MIRROR study. In: *Neurology* 90 (20), e1805-e1814. DOI: 10.1212/WNL.0000000000005516.

Basaran, Berke Doga; Matthews, Paul M.; Bai, Wenjia (2022): New lesion segmentation for multiple sclerosis brain images with imaging and lesion-aware augmentation. In: *Frontiers in neuroscience* 16. DOI: 10.3389/fnins.2022.1007453.

Bayas, A.; Berthele, A.; Hemmer, B.; Warnke, C.; Wildemann, B. (2021): Controversy on the treatment of multiple sclerosis and related disorders: positional statement of the expert panel in charge of the 2021 DGN Guideline on diagnosis and treatment of multiple sclerosis, neuromyelitis optica spectrum diseases and MOG-IgG-associated disorders. In: *Neurological research and practice* 3 (1), S. 45. DOI: 10.1186/s42466-021-00139-8.

Berthele, Achim; Hemmer, B. (2021): S2k-Leitlinie: Diagnose und Therapie der Multiplen Sklerose, Neuromyelitis-optica-Spektrum-Erkrankungen und MOG-IgG-assoziierten Erkrankungen. In: *DGNeurologie* 4 (4), S. 251–275. DOI: 10.1007/s42451-021-00334-6.

Bevan, Carolyn J.; Cree, Bruce A. C. (2014): Disease activity free status: a new end point for a new era in multiple sclerosis clinical research? In: *JAMA neurology* 71 (3), S. 269–270. DOI: 10.1001/jamaneurol.2013.5486.

Bonacchi, Raffaello; Filippi, Massimo; Rocca, Maria A. (2022): Role of artificial intelligence in MS clinical practice. In: *NeuroImage. Clinical* 35. DOI: 10.1016/j.nicl.2022.103065.

Bozsik, Bence; Tóth, Eszter; Polyák, Ilona; Kerekes, Fanni; Szabó, Nikoletta; Bencsik, Krisztina et al. (2022): Reproducibility of Lesion Count in Various Subregions on MRI Scans in Multiple Sclerosis. In: *Frontiers in neurology* 13. DOI: 10.3389/fneur.2022.843377.

Brønnum-Hansen, Henrik; Koch-Henriksen, Nils; Stenager, Egon (2004): Trends in survival and cause of death in Danish patients with multiple sclerosis. In: *Brain : a journal of neurology* 127 (Pt 4), S. 844–850. DOI: 10.1093/brain/awh104.

Brosch, Tom; Tang, Lisa Y. W.; Youngjin, Yoo; Li, David K. B.; Traboulsee, Anthony; Tam, Roger (2022): Deep 3D Convolutional Encoder Networks With Shortcuts for Multiscale Feature Integration Applied to Multiple Sclerosis Lesion Segmentation. In: *IEEE transactions on medical imaging* 35 (5). DOI: 10.1109/TMI.2016.2528821.

Browne, Paul; Chandraratna, Dhia; Angood, Ceri; Tremlett, Helen; Baker, Chris; Taylor, Bruce V.; Thompson, Alan J. (2014): Atlas of Multiple Sclerosis 2013: A growing global problem with widespread inequity. In: *Neurology* 83 (11), S. 1022–1024. DOI: 10.1212/WNL.0000000000000768.

Brownlee, Wallace J.; Hardy, Todd A.; Fazekas, Franz; Miller, David H. (2017): Diagnosis of multiple sclerosis: progress and challenges. In: *Lancet (London, England)* 389 (10076), S. 1336–1346. DOI: 10.1016/S0140-6736(16)30959-X.

Burt, Richard K.; Balabanov, Roumen; Burman, Joachim; Sharrack, Basil; Snowden, John A.; Oliveira, Maria Carolina et al. (2019): Effect of Nonmyeloablative Hematopoietic Stem Cell Transplantation vs Continued Disease-Modifying Therapy on Disease Progression in Patients With Relapsing-Remitting Multiple Sclerosis: A Randomized Clinical Trial. In: *JAMA* 321 (2), S. 165–174. DOI: 10.1001/jama.2018.18743.

Cabezas, M.; Corral, J. F.; Oliver, A.; Díez, Y.; Tintoré, M.; Auger, C. et al. (2016): Improved Automatic Detection of New T2 Lesions in Multiple Sclerosis Using Deformation Fields. In: *AJNR. American journal of neuroradiology* 37 (10), S. 1816–1823. DOI: 10.3174/ajnr.A4829.

Calabrese, Massimiliano; Atzori, Matteo; Bernardi, Valentina; Morra, Aldo; Romualdi, Chiara; Rinaldi, Luciano et al. (2007): Cortical atrophy is relevant in multiple sclerosis at clinical onset. In: *Journal of neurology* 254 (9), S. 1212–1220. DOI: 10.1007/s00415-006-0503-6.

Calabresi, Peter A.; Radue, Ernst-Wilhelm; Goodin, Douglas; Jeffery, Douglas; Rammohan, Kottil W.; Reder, Anthony T. et al. (2014): Safety and efficacy of fingolimod in patients with relapsing-remitting multiple sclerosis (FREEDOMS II): a double-blind, randomised, placebo-controlled, phase 3 trial. In: *The Lancet. Neurology* 13 (6), S. 545–556. DOI: 10.1016/S1474-4422(14)70049-3.

Calcagno, Giuseppe; Staiano, Antonino; Fortunato, Giuliana; Brescia-Morra, Vincenzo; Salvatore, Elena; Liguori, Rosario et al. (2010): A multilayer perceptron neural network-based approach for the identification of responsiveness to interferon therapy in multiple sclerosis patients. In: *Information Sciences* 180 (21), S. 4153–4163. DOI: 10.1016/j.ins.2010.07.004.

Carass, Aaron; Roy, Snehashis; Jog, Amod; Cuzzocreo, Jennifer L.; Magrath, Elizabeth; Gherman, Adrian et al. (2017): Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. In: *NeuroImage* 148, S. 77–102. DOI: 10.1016/j.neuroimage.2016.12.064.

Carlström, Karl E.; Ewing, Ewoud; Granqvist, Mathias; Gyllenberg, Alexandra; Aeinehband, Shahin; Enoksson, Sara Lind et al. (2019): Therapeutic efficacy of dimethyl fumarate in relapsing-remitting multiple sclerosis associates with ROS

pathway in monocytes. In: *Nature communications* 10 (1), S. 3081. DOI: 10.1038/s41467-019-11139-3.

Chung, Karen K.; Altmann, Daniel; Barkhof, Frederik; Miszkiel, Katherine; Brex, Peter A.; O'Riordan, Jonathan et al. (2020): A 30-Year Clinical and Magnetic Resonance Imaging Observational Study of Multiple Sclerosis and Clinically Isolated Syndromes. In: *Annals of neurology* 87 (1), S. 63–74. DOI: 10.1002/ana.25637.

Cobo-Calvo, Álvaro; d'Indy, Hyacintha; Ruiz, Anne; Collongues, Nicolas; Kremer, Laurent; Durand-Dubief, Françoise et al. (2020): Frequency of myelin oligodendrocyte glycoprotein antibody in multiple sclerosis: A multicenter cross-sectional study. In: *Neurology(R) neuroimmunology & neuroinflammation* 7 (2). DOI: 10.1212/NXI.0000000000000649.

Cohen, Jacob (1960): A Coefficient of Agreement for Nominal Scales.

Cohen, Jeffrey A.; Barkhof, Frederik; Comi, Giancarlo; Hartung, Hans-Peter; Khatri, Bhupendra O.; Montalban, Xavier et al. (2010): Oral fingolimod or intramuscular interferon for relapsing multiple sclerosis. In: *The New England journal of medicine* 362 (5), S. 402–415. DOI: 10.1056/NEJMoa0907839.

Cohen, Jeffrey A.; Comi, Giancarlo; Selmaj, Krzysztof W.; Bar-Or, Amit; Arnold, Douglas L.; Steinman, Lawrence et al. (2019): Safety and efficacy of ozanimod versus interferon beta-1a in relapsing multiple sclerosis (RADIANCE): a multicentre, randomised, 24-month, phase 3 trial. In: *The Lancet. Neurology* 18 (11), S. 1021–1033. DOI: 10.1016/S1474-4422(19)30238-8.

Combès, Benoit; Kerbrat, Anne; Pasquier, Guillaume; Commowick, Olivier; Le Bon, Brandon; Galassi, Francesca et al. (2021): A Clinically-Compatible Workflow for Computer-Aided Assessment of Brain Disease Activity in Multiple Sclerosis Patients. In: *Frontiers in medicine* 8, S. 740248. DOI: 10.3389/fmed.2021.740248.

Comi, Giancarlo; Kappos, Ludwig; Selmaj, Krzysztof W.; Bar-Or, Amit; Arnold, Douglas L.; Steinman, Lawrence et al. (2019): Safety and efficacy of ozanimod versus interferon beta-1a in relapsing multiple sclerosis (SUNBEAM): a

multicentre, randomised, minimum 12-month, phase 3 trial. In: *The Lancet. Neurology* 18 (11), S. 1009–1020. DOI: 10.1016/S1474-4422(19)30239-X.

Commowick, Olivier; Istace, Audrey; Kain, Michaël; Laurent, Baptiste; Leray, Florent; Simon, Mathieu et al. (2018): Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure. In: *Scientific reports* 8 (1), S. 13650. DOI: 10.1038/s41598-018-31911-7.

Commowick, Olivier; Kain, Michaël; Casey, Romain; Ameli, Roxana; Ferré, Jean-Christophe; Kerbrat, Anne et al. (2021): Multiple sclerosis lesions segmentation from multiple experts: The MICCAI 2016 challenge dataset. In: *NeuroImage* 244. DOI: 10.1016/j.neuroimage.2021.118589.

Compston, Alastair; Coles, Alasdair (2008): Multiple sclerosis 372 (9648), S. 1502–1517. DOI: 10.1016/S0140-6736(08)61620-7.

Cross, Helen; Sabiq, Farahna; Ackermans, Nathalie; Mattar, Andrew; Au, Shelly; Woodhall, Mark et al. (2020): Myelin Oligodendrocyte Glycoprotein (MOG) Antibody Positive Patients in a Multi-Ethnic Canadian Cohort. In: *Frontiers in neurology* 11, S. 525933. DOI: 10.3389/fneur.2020.525933.

Dahan, A.; Pereira, R.; Malpas, C. B.; Kalincik, T.; Gaillard, F. (2019): PACS Integration of Semiautomated Imaging Software Improves Day-to-Day MS Disease Activity Detection. In: *AJNR. American journal of neuroradiology* 40 (10), S. 1624–1629. DOI: 10.3174/ajnr.A6195.

Dahan, Ariel; Wang, Wayland; Gaillard, Frank (2018): Computer-Aided Detection Can Bridge the Skill Gap in Multiple Sclerosis Monitoring. In: *Journal of the American College of Radiology : JACR* 15 (1 Pt A), S. 93–96. DOI: 10.1016/j.jacr.2017.06.030.

Danelakis, Antonios; Theoharis, Theoharis; Verganelakis, Dimitrios A. (2018): Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging 2018.

Dobson, Ruth; Ramagopalan, Sreeram; Davis, Angharad; Giovannoni, Gavin (2013): Cerebrospinal fluid oligoclonal bands in multiple sclerosis and clinically isolated syndromes: a meta-analysis of prevalence, prognosis and effect of

latitude. In: *Journal of neurology, neurosurgery, and psychiatry* 84 (8), S. 909–914. DOI: 10.1136/jnnp-2012-304695.

Dyment, David A.; Ebers, George C.; Sadovnick, A. Dessa (2004): Genetics of multiple sclerosis. In: *The Lancet. Neurology* 3 (2), S. 104–110. DOI: 10.1016/s1474-4422(03)00663-x.

Ebers, George C. (2008): Environmental factors and multiple sclerosis. In: *The Lancet. Neurology* 7 (3), S. 268–277. DOI: 10.1016/S1474-4422(08)70042-5.

Egger, Christine; Opfer, Roland; Wang, Chenyu; Kepp, Timo; Sormani, Maria Pia; Spies, Lothar et al. (2017): MRI FLAIR lesion segmentation in multiple sclerosis: Does automated segmentation hold up with manual annotation? In: *NeuroImage. Clinical* 13, S. 264–270. DOI: 10.1016/j.nicl.2016.11.020.

Eichinger, Paul; Zimmer, Claus; Wiestler, Benedikt (2020): KI in der Radiologie: Wo stehen wir in der MS-Bildgebung? In: *RoFo : Fortschritte auf dem Gebiete der Rontgenstrahlen und der Nuklearmedizin* 192 (9), S. 847–853. DOI: 10.1055/a-1167-8402.

Eitel, Fabian; Soehler, Emily; Bellmann-Strobl, Judith; Brandt, Alexander U.; Ruprecht, Klemens; Giess, René M. et al. (2019): Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation. In: *NeuroImage. Clinical* 24, S. 102003. DOI: 10.1016/j.nicl.2019.102003.

European Society of Radiology 2009 (2010): The future role of radiology in healthcare. In: *Insights into imaging* 1 (1), S. 2–11. DOI: 10.1007/s13244-009-0007-x.

Fahrbach, Kyle; Huelin, Rachel; Martin, Amber L.; Kim, Edward; Dastani, Homa B.; Rao, Stephen; Malhotra, Manoj (2013): Relating relapse and T2 lesion changes to disability progression in multiple sclerosis: a systematic literature review and regression analysis. In: *BMC neurology* 13, S. 180. DOI: 10.1186/1471-2377-13-180.

Fartaria, Mário João; Bonnier, Guillaume; Roche, Alexis; Kober, Tobias; Meuli, Reto; Rotzinger, David et al. (2016): Automated detection of white matter and

cortical lesions in early stages of multiple sclerosis. In: *Journal of magnetic resonance imaging : JMRI* 43 (6), S. 1445–1454. DOI: 10.1002/jmri.25095.

Fazekas, F.; Barkhof, F.; Filippi, M.; Grossman, R. I.; Li, D. K.; McDonald, W. I. et al. (1999): The contribution of magnetic resonance imaging to the diagnosis of multiple sclerosis. In: *Neurology* 53 (3), S. 448–456. DOI: 10.1212/wnl.53.3.448.

Fedorov, Andriy; Beichel, Reinhard; Kalpathy-Cramer, Jayashree; Finet, Julien; Fillion-Robin, Jean-Christophe; Pujol, Sonia et al. (2012): 3D Slicer as an image computing platform for the Quantitative Imaging Network. In: *Magnetic resonance imaging* 30 (9), S. 1323–1341. DOI: 10.1016/j.mri.2012.05.001.

Filippi, M.; Horsfield, M. A.; Adèr, H. J.; Barkhof, F.; Bruzzi, P.; Evans, A. et al. (1998): Guidelines for using quantitative measures of brain magnetic resonance imaging abnormalities in monitoring the treatment of multiple sclerosis. In: *Annals of neurology* 43 (4), S. 499–506. DOI: 10.1002/ana.410430414.

Filippi, M.; Horsfield, M. A.; Campi, A.; Mammi, S.; Pereira, C.; Comi, G. (1995): Resolution-dependent estimates of lesion volumes in magnetic resonance imaging studies of the brain in multiple sclerosis. In: *Annals of neurology* 38 (5), S. 749–754. DOI: 10.1002/ana.410380509.

Filippi, M.; Rocca, M. A. (2010): Dirty-appearing white matter: a disregarded entity in multiple sclerosis. In: *AJNR. American journal of neuroradiology* 31 (3), S. 390–391. DOI: 10.3174/ajnr.A1834.

Filippi, Massimo; Bar-Or, Amit; Piehl, Fredrik; Preziosa, Paolo; Solari, Alessandra; Vukusic, Sandra; Rocca, Maria A. (2018): Multiple sclerosis. In: *Nature reviews. Disease primers* 4 (1), S. 43. DOI: 10.1038/s41572-018-0041-4.

Filippi, Massimo; Preziosa, Paolo; Banwell, Brenda L.; Barkhof, Frederik; Ciccarelli, Olga; Stefano, Nicola de et al. (2019): Assessment of lesions on magnetic resonance imaging in multiple sclerosis: practical guidelines. In: *Brain : a journal of neurology* 142 (7), S. 1858–1875. DOI: 10.1093/brain/awz144.

Filippi, Massimo; Rocca, Maria A.; Ciccarelli, Olga; Stefano, Nicola de; Evangelou, Nikos; Kappos, Ludwig et al. (2016): MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines. In: *The Lancet. Neurology* 15 (3), S. 292–303. DOI: 10.1016/S1474-4422(15)00393-2.

Fisniku, L. K.; Brex, P. A.; Altmann, D. R.; Miszkiel, K. A.; Benton, C. E.; Lanyon, R. et al. (2008): Disability and T2 MRI lesions: a 20-year follow-up of patients with relapse onset of multiple sclerosis. In: *Brain : a journal of neurology* 131 (Pt 3), S. 808–817. DOI: 10.1093/brain/awm329.

Flachenecker, Peter; Kobelt, Gisela; Berg, Jenny; Capsa, Daniela; Gannedahl, Mia (2017): New insights into the burden and costs of multiple sclerosis in Europe: Results for Germany. In: *Multiple sclerosis (Houndmills, Basingstoke, England)* 23 (2_suppl), S. 78–90. DOI: 10.1177/1352458517708141.

Gabr, Refaat E.; Coronado, Ivan; Robinson, Melvin; Sujit, Sheeba J.; Datta, Sushmita; Sun, Xiaojun et al. (2020): Brain and lesion segmentation in multiple sclerosis using fully convolutional neural networks: A large-scale study. In: *Multiple sclerosis (Houndmills, Basingstoke, England)* 26 (10), S. 1217–1226. DOI: 10.1177/1352458519856843.

Galletto Pregliasco, A.; Collin, A.; Guéguen, A.; Metten, M. A.; Aboab, J.; Deschamps, R. et al. (2018): Improved Detection of New MS Lesions during Follow-Up Using an Automated MR Coregistration-Fusion Method. In: *AJNR. American journal of neuroradiology* 39 (7), S. 1226–1232. DOI: 10.3174/ajnr.A5690.

García-Lorenzo, Daniel; Francis, Simon; Narayanan, Sridar; Arnold, Douglas L.; Collins, D. Louis (2013): Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. In: *Medical image analysis* 17 (1), S. 1–18. DOI: 10.1016/j.media.2012.09.004.

Gasperini, Claudio; Ruggieri, Serena; Mancinelli, Chiara Rosa; Pozzilli, Carlo (2013): Advances in the treatment of relapsing-remitting multiple sclerosis - critical appraisal of fingolimod. In: *Therapeutics and clinical risk management* 9, S. 73–85. DOI: 10.2147/TCRM.S17426.

Geremia, Ezequiel; Clatz, Olivier; Menze, Bjoern H.; Konukoglu, Ender; Criminisi, Antonio; Ayache, Nicholas (2011): Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. In: *NeuroImage* 57 (2), S. 378–390. DOI: 10.1016/j.neuroimage.2011.03.080.

Ghasemi, Nazem; Razavi, Shahnaz; Nikzad, Elham (2017): Multiple Sclerosis: Pathogenesis, Symptoms, Diagnoses and Cell-Based Therapy. In: *Cell journal* 19 (1), S. 1–10. DOI: 10.22074/cellj.2016.4867.

Giovannoni, Gavin; Butzkueven, Helmut; Dhib-Jalbut, Suhayl; Hobart, Jeremy; Kobelt, Gisela; Pepper, George et al. (2016): Brain health: time matters in multiple sclerosis. In: *Multiple sclerosis and related disorders* 9 Suppl 1, S5-S48. DOI: 10.1016/j.msard.2016.07.003.

Giovannoni, Gavin; Soelberg Sorensen, Per; Cook, Stuart; Rammohan, Kottil; Rieckmann, Peter; Comi, Giancarlo et al. (2018): Safety and efficacy of cladribine tablets in patients with relapsing-remitting multiple sclerosis: Results from the randomized extension trial of the CLARITY study. In: *Multiple sclerosis (Houndmills, Basingstoke, England)* 24 (12), S. 1594–1604. DOI: 10.1177/1352458517727603.

Giovannoni, Gavin; Turner, Benjamin; Gnanapavan, Sharmilee; Offiah, Curtis; Schmierer, Klaus; Marta, Monica (2015): Is it time to target no evident disease activity (NEDA) in multiple sclerosis? In: *Multiple sclerosis and related disorders* 4 (4), S. 329–333. DOI: 10.1016/j.msard.2015.04.006.

Gold, Ralf; Arnold, Douglas L.; Bar-Or, Amit; Fox, Robert J.; Kappos, Ludwig; Mokliatchouk, Oksana et al. (2022): Long-term safety and efficacy of dimethyl fumarate for up to 13 years in patients with relapsing-remitting multiple sclerosis: Final ENDORSE study results. In: *Multiple sclerosis (Houndmills, Basingstoke, England)* 28 (5), S. 801–816. DOI: 10.1177/13524585211037909.

Goldenberg, Marvin M. (2012): Multiple sclerosis review. In: *P & T : a peer-reviewed journal for formulary management* 37 (3), S. 175–184.

Goodin, Douglas S.; Traboulsee, Anthony; Knappertz, Volker; Reder, Anthony T.; Li, David; Langdon, Dawn et al. (2012): Relationship between early clinical

characteristics and long term disability outcomes: 16 year cohort study (follow-up) of the pivotal interferon β-1b trial in multiple sclerosis. In: *Journal of neurology, neurosurgery, and psychiatry* 83 (3), S. 282–287. DOI: 10.1136/jnnp-2011-301178.

Grahl, S.; Pongratz, V.; Schmidt, P.; Engl, C.; Bussas, M.; Radetz, A. et al. (2019): Evidence for a white matter lesion size threshold to support the diagnosis of relapsing remitting multiple sclerosis. In: *Multiple sclerosis and related disorders* 29, S. 124–129. DOI: 10.1016/j.msard.2019.01.042.

Gramsch, Carolin; Nensa, Felix; Kastrup, Oliver; Maderwald, Stefan; Deuschl, Cornelius; Ringelstein, Adrian et al. (2015): Diagnostic value of 3D fluid attenuated inversion recovery sequence in multiple sclerosis. In: *Acta radiologica (Stockholm, Sweden : 1987)* 56 (5), S. 622–627. DOI: 10.1177/0284185114534413.

Grauer, O.; Offenhäusser, M.; Schmidt, J.; Toyka, K. V.; Gold, R. (2001): Glucocorticosteroid therapy in optic neuritis and multiple sclerosis. Evidence from clinical studies and practical recommendations. In: *Der Nervenarzt* 72 (8), S. 577–589. DOI: 10.1007/s001150170057.

Grimaud, J.; Lai, M.; Thorpe, J.; Adeleine, P.; Wang, L.; Barker, G. J. et al. (1996): Quantification of MRI lesion load in multiple sclerosis: a comparison of three computer-assisted techniques. In: *Magnetic resonance imaging* 14 (5), S. 495–505. DOI: 10.1016/0730-725x(96)00018-5.

Halilovic, Emina Alimanovic; Alimanovic, Ilda; Suljic, Enra; Hassan, Nabil Al (2014): Optic neuritis as first clinical manifestations the multiple sclerosis. In: *Materia socio-medica* 26 (4), S. 246–248. DOI: 10.5455/msm.2014.246-2481.

Hamet, Pavel; Tremblay, Johanne (2017): Artificial intelligence in medicine. In: *Metabolism: clinical and experimental* 69S, S36-S40. DOI: 10.1016/j.metabol.2017.01.011.

Hauser, Stephen L.; Bar-Or, Amit; Cohen, Jeffrey A.; Comi, Giancarlo; Correale, Jorge; Coyle, Patricia K. et al. (2020a): Ofatumumab versus Teriflunomide in

Multiple Sclerosis. In: *The New England journal of medicine* 383 (6), S. 546–557. DOI: 10.1056/NEJMoa1917246.

Hauser, Stephen L.; Kappos, Ludwig; Arnold, Douglas L.; Bar-Or, Amit; Brochet, Bruno; Naismith, Robert T. et al. (2020b): Five years of ocrelizumab in relapsing multiple sclerosis: OPERA studies open-label extension. In: *Neurology* 95 (13), e1854-e1867. DOI: 10.1212/WNL.0000000000010376.

Hawkes, C. H.; Macgregor, A. J. (2009): Twin studies and the heritability of MS: a conclusion. In: *Multiple sclerosis (Houndmills, Basingstoke, England)* 15 (6), S. 661–667. DOI: 10.1177/1352458509104592.

Hemmer B. et al. (2023): Diagnose und Therapie der Multiplen Sklerose, Neuromyelitis-optica-Spektrum-Erkrankungen und MOG-IgG-assoziierten Erkrankungen. In: *Deutsche Gesellschaft für Neurologie (Hrsg.), Leitlinien für Diagnostik und Therapie in der Neurologie.*

Hervás-García, J. V.; Ramió-Torrentà, L.; Brieva-Ruiz, L.; Batllé-Nadal, J.; Moral, E.; Blanco, Y. et al. (2019): Comparison of two high doses of oral methylprednisolone for multiple sclerosis relapses: a pilot, multicentre, randomized, double-blind, non-inferiority trial. In: *European journal of neurology* 26 (3), S. 525–532. DOI: 10.1111/ene.13851.

Hindsholm, Amalie Monberg; Cramer, Stig Præstekjær; Simonsen, Helle Juhl; Frederiksen, Jette Lautrup; Andersen, Flemming; Højgaard, Liselotte et al. (2022): Assessment of Artificial Intelligence Automatic Multiple Sclerosis Lesion Delineation Tool for Clinical Use. In: *Clinical neuroradiology* 32 (3), S. 643–653. DOI: 10.1007/s00062-021-01089-z.

Hitziger, Sebastian; Ling, Wen Xin; Fritz, Thomas; D'Albis, Tiziano; Lemke, Andreas; Grilo, Joana (2022): Triplanar U-Net with lesion-wise voting for the segmentation of new lesions on longitudinal MRI studies. In: *Frontiers in neuroscience* 16. DOI: 10.3389/fnins.2022.964250.

Hohlfeld, Reinhard (2010): 'Gimme five': future challenges in multiple sclerosis. ECTRIMS Lecture 2009. In: *Multiple sclerosis (Houndmills, Basingstoke, England)* 16 (1), S. 3–14. DOI: 10.1177/1352458509357355.

Hohlfeld, Reinhard; Dornmair, Klaus; Meinl, Edgar; Wekerle, Hartmut (2016): The search for the target antigens of multiple sclerosis, part 1: autoreactive CD4+ T lymphocytes as pathogenic effectors and therapeutic targets. In: *The Lancet. Neurology* 15 (2), S. 198–209. DOI: 10.1016/S1474-4422(15)00334-8.

Jakob Holstiege; Annika Steffen; Benjamin Goffrier; Jörg Bätzing (2017): Epidemiologie der Multiplen Sklerose – Eine populationsbasierte deutschlandweite Studie.

Jesson, Andrew; Arbel, Tal: Hierarchical MRF and random forest segmentation of MS lesions and healthy tissues in brain MRI.

Johnson, K. P.; Brooks, B. R.; Ford, C. C.; Goodman, A.; Guarnaccia, J.; Lisak, R. P. et al. (2000): Sustained clinical benefits of glatiramer acetate in relapsing multiple sclerosis patients observed for 6 years. Copolymer 1 Multiple Sclerosis Study Group. In: *Multiple sclerosis (Houndmills, Basingstoke, England)* 6 (4), S. 255–266. DOI: 10.1177/135245850000600407.

Kapoor, Raju; Ho, Pei-Ran; Campbell, Nolan; Chang, Ih; Deykin, Aaron; Forrestal, Fiona et al. (2018): Effect of natalizumab on disease progression in secondary progressive multiple sclerosis (ASCEND): a phase 3, randomised, double-blind, placebo-controlled trial with an open-label extension. In: *The Lancet. Neurology* 17 (5), S. 405–415. DOI: 10.1016/S1474-4422(18)30069-3.

Kappos, Ludwig; Bar-Or, Amit; Cree, Bruce A. C.; Fox, Robert J.; Giovannoni, Gavin; Gold, Ralf et al. (2018): Siponimod versus placebo in secondary progressive multiple sclerosis (EXPAND): a double-blind, randomised, phase 3 study. In: *Lancet (London, England)* 391 (10127), S. 1263–1273. DOI: 10.1016/S0140-6736(18)30475-6.

Kappos, Ludwig; Fox, Robert J.; Burcklen, Michel; Freedman, Mark S.; Havrdová, Eva K.; Hennessy, Brian et al. (2021): Ponesimod Compared With Teriflunomide in Patients With Relapsing Multiple Sclerosis in the Active-Comparator Phase 3 OPTIMUM Study: A Randomized Clinical Trial. In: *JAMA neurology* 78 (5), S. 558–567. DOI: 10.1001/jamaneurol.2021.0405.

Kappos, Ludwig; Gold, Ralf; Miller, David H.; Macmanus, David G.; Havrdova, Eva; Limmroth, Volker et al. (2008): Efficacy and safety of oral fumarate in patients with relapsing-remitting multiple sclerosis: a multicentre, randomised, double-blind, placebo-controlled phase IIb study. In: *Lancet (London, England)* 372 (9648), S. 1463–1472. DOI: 10.1016/S0140-6736(08)61619-0.

Kappos, Ludwig; Li, David; Calabresi, Peter A.; O'Connor, Paul; Bar-Or, Amit; Barkhof, Frederik et al. (2011): Ocrelizumab in relapsing-remitting multiple sclerosis: a phase 2, randomised, placebo-controlled, multicentre trial. In: *Lancet (London, England)* 378 (9805), S. 1779–1787. DOI: 10.1016/S0140-6736(11)61649-8.

Kappos, Ludwig; Stefano, Nicola de; Freedman, Mark S.; Cree, Bruce Ac; Radue, Ernst-Wilhelm; Sprenger, Till et al. (2016): Inclusion of brain volume loss in a revised measure of 'no evidence of disease activity' (NEDA-4) in relapsing-remitting multiple sclerosis. In: *Multiple sclerosis (Houndmills, Basingstoke, England)* 22 (10), S. 1297–1305. DOI: 10.1177/1352458515616701.

Khan, Omar; Rieckmann, Peter; Boyko, Alexey; Selmaj, Krzysztof; Zivadinov, Robert (2013): Three times weekly glatiramer acetate in relapsing-remitting multiple sclerosis. In: *Annals of neurology* 73 (6), S. 705–713. DOI: 10.1002/ana.23938.

Kicky G. van Leeuwen; Frederick J. A. Meijer; Steven Schalekamp; Matthieu J. C. M. Rutten; Ewoud J. van Dijk; Bram van Ginneken et al.: Cost-effectiveness of artificial intelligence aided vessel occlusion detection in acute stroke: an early health technology assessment.

Krüger, Julia; Opfer, Roland; Gessert, Nils; Ostwaldt, Ann-Christin; Manogaran, Praveena; Kitzler, Hagen H. et al. (2020): Fully automated longitudinal segmentation of new or enlarged multiple sclerosis lesions using 3D convolutional neural networks. In: *NeuroImage. Clinical* 28, S. 102445. DOI: 10.1016/j.nicl.2020.102445.

Krüger, Julia; Ostwaldt, Ann-Christin; Spies, Lothar; Geisler, Benjamin; Schlaefer, Alexander; Kitzler, Hagen H. et al. (2022): Infratentorial lesions in

multiple sclerosis patients: intra- and inter-rater variability in comparison to a fully automated segmentation using 3D convolutional neural networks. In: *European radiology* 32 (4), S. 2798–2809. DOI: 10.1007/s00330-021-08329-3.

Krummenauer, F.; Kauczor, H. U. (2002): Fallzahlplanung in referenzkontrollierten Diagnosestudien. In: *RoFo : Fortschritte auf dem Gebiete der Rontgenstrahlen und der Nuklearmedizin* 174 (11), S. 1438–1444. DOI: 10.1055/s-2002-35346.

Landis, J. R.; Koch, G. G. (1977): The measurement of observer agreement for categorical data. In: *Biometrics* 33 (1), S. 159–174.

Lebrun-Frenay, Christine; Kantarci, Orhun; Siva, Aksel; Sormani, Maria P.; Pelletier, Daniel; Okuda, Darin T. (2020): Radiologically Isolated Syndrome: 10-Year Risk Estimate of a Clinical Event. In: *Annals of neurology* 88 (2), S. 407–417. DOI: 10.1002/ana.25799.

LeCun, Yann; Bengio, Yoshua; Hinton, Geoffrey (2015): Deep learning. In: *Nature* 521 (7553), S. 436–444. DOI: 10.1038/nature14539.

Leocani, Letizia; Comi, Giancarlo (2008): Neurophysiological markers. In: *Neurological sciences : official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology* 29 Suppl 2, S218-21. DOI: 10.1007/s10072-008-0942-1.

Levin, Lynn I.; Munger, Kassandra L.; Rubertone, Mark V.; Peck, Charles A.; Lennette, Evelyne T.; Spiegelman, Donna; Ascherio, Alberto (2005): Temporal relationship between elevation of epstein-barr virus antibody titers and initial onset of neurological symptoms in multiple sclerosis. In: *JAMA* 293 (20), S. 2496–2500. DOI: 10.1001/jama.293.20.2496.

Liu, Rongzeng; Du, Shushu; Zhao, Lili; Jain, Sahil; Sahay, Kritika; Rizvanov, Albert et al. (2022): Autoreactive lymphocytes in multiple sclerosis: Pathogenesis and treatment target. In: *Frontiers in immunology* 13, S. 996469. DOI: 10.3389/fimmu.2022.996469.

Lublin, F. D.; Reingold, S. C. (1996): Defining the clinical course of multiple sclerosis: results of an international survey. National Multiple Sclerosis Society

(USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis. In: *Neurology* 46 (4), S. 907–911. DOI: 10.1212/wnl.46.4.907.

Lundervold, Alexander Selvikvåg; Lundervold, Arvid (2019): An overview of deep learning in medical imaging focusing on MRI. In: *Zeitschrift fur medizinische Physik* 29 (2), S. 102–127. DOI: 10.1016/j.zemedi.2018.11.002.

Lycklama à Nijeholt, G. J. (2005): Reduction of brain volume in MS. MRI and pathology findings. In: *Journal of the neurological sciences* 233 (1-2), S. 199–202. DOI: 10.1016/j.jns.2005.03.016.

Ma, Yang; Zhang, Chaoyi; Cabezas, Mariano; Song, Yang; Tang, Zihao; Liu, Dongnan et al. (2022): Multiple Sclerosis Lesion Analysis in Brain Magnetic Resonance Images: Techniques and Clinical Applications. In: *IEEE journal of biomedical and health informatics* 26 (6), S. 2680–2692. DOI: 10.1109/JBHI.2022.3151741.

McFarland, H. F.; Frank, J. A.; Albert, P. S.; Smith, M. E.; Martin, R.; Harris, J. O. et al. (1992): Using gadolinium-enhanced magnetic resonance imaging lesions to monitor disease activity in multiple sclerosis. In: *Annals of neurology* 32 (6), S. 758–766. DOI: 10.1002/ana.410320609.

Metz, Luanne M.; Li, David K. B.; Traboulsee, Anthony L.; Duquette, Pierre; Eliasziw, Misha; Cerchiaro, Graziela et al. (2017): Trial of Minocycline in a Clinically Isolated Syndrome of Multiple Sclerosis. In: *The New England journal of medicine* 376 (22), S. 2122–2133. DOI: 10.1056/NEJMoa1608889.

Miller, Aaron E.; Olsson, Tomas P.; Wolinsky, Jerry S.; Comi, Giancarlo; Kappos, Ludwig; Hu, Xueqiang et al. (2020): Long-term safety and efficacy of teriflunomide in patients with relapsing multiple sclerosis: Results from the TOWER extension study. In: *Multiple sclerosis and related disorders* 46, S. 102438. DOI: 10.1016/j.msard.2020.102438.

Milo, Ron; Kahana, Esther (2010): Multiple sclerosis: geoepidemiology, genetics and the environment. In: *Autoimmunity reviews* 9 (5), A387-94. DOI: 10.1016/j.autrev.2009.11.010.

Molyneux, P. D.; Miller, D. H.; Filippi, M.; Yousry, T. A.; Radü, E. W.; Adèr, H. J.; Barkhof, F. (1999): Visual analysis of serial T2-weighted MRI in multiple sclerosis: intra- and interobserver reproducibility. In: *Neuroradiology* 41 (12), S. 882–888. DOI: 10.1007/s002340050860.

Montalban, Xavier; Leist, Thomas P.; Cohen, Bruce A.; Moses, Harold; Campbell, Jackie; Hicking, Christine; Dangond, Fernando (2018): Cladribine tablets added to IFN-β in active relapsing MS: The ONWARD study. In: *Neurology(R) neuroimmunology & neuroinflammation* 5 (5), e477. DOI: 10.1212/NXI.0000000000000477.

Mortazavi, Daryoush; Kouzani, Abbas Z.; Soltanian-Zadeh, Hamid (2012): Segmentation of multiple sclerosis lesions in MR images: a review. In: *Neuroradiology* 54 (4), S. 299–320. DOI: 10.1007/s00234-011-0886-7.

Muthukrishnan, Nikesh; Maleki, Farhad; Ovens, Katie; Reinhold, Caroline; Forghani, Behzad; Forghani, Reza (2020): Brief History of Artificial Intelligence. In: *Neuroimaging clinics of North America* 30 (4), S. 393–399. DOI: 10.1016/j.nic.2020.07.004.

Naji, Yahya; Mahdaoui, Mohamed; Klevor, Raymond; Kissani, Najib (2023): Artificial Intelligence and Multiple Sclerosis: Up-to-Date Review. In: *Cureus* 15 (9), e45412. DOI: 10.7759/cureus.45412.

Neema, M.; Guss, Z. D.; Stankiewicz, J. M.; Arora, A.; Healy, B. C.; Bakshi, R. (2009): Normal findings on brain fluid-attenuated inversion recovery MR images at 3T. In: *AJNR. American journal of neuroradiology* 30 (5), S. 911–916. DOI: 10.3174/ajnr.A1514.

Noseworthy, J. H.; Lucchinetti, C.; Rodriguez, M.; Weinshenker, B. G. (2000): Multiple sclerosis. In: *The New England journal of medicine* 343 (13), S. 938–952. DOI: 10.1056/NEJM200009283431307.

O'Connor, Paul; Comi, Giancarlo; Freedman, Mark S.; Miller, Aaron E.; Kappos, Ludwig; Bouchard, Jean-Pierre et al. (2016): Long-term safety and efficacy of teriflunomide: Nine-year follow-up of the randomized TEMSO study. In: *Neurology* 86 (10), S. 920–930. DOI: 10.1212/WNL.0000000000002441.

O'Connor, Paul; Goodman, Andrew; Kappos, Ludwig; Lublin, Fred; Polman, Chris; Rudick, Richard A. et al. (2014): Long-term safety and effectiveness of natalizumab redosing and treatment in the STRATA MS Study. In: *Neurology* 83 (1), S. 78–86. DOI: 10.1212/WNL.0000000000000541.

Okuda, D. T.; Mowry, E. M.; Beheshtian, A.; Waubant, E.; Baranzini, S. E.; Goodin, D. S. et al. (2009): Incidental MRI anomalies suggestive of multiple sclerosis: the radiologically isolated syndrome. In: *Neurology* 72 (9), S. 800–805. DOI: 10.1212/01.wnl.0000335764.14513.1a.

Olsson, Tomas; Boster, Aaron; Fernández, Óscar; Freedman, Mark S.; Pozzilli, Carlo; Bach, Doris et al. (2014): Oral ponesimod in relapsing-remitting multiple sclerosis: a randomised phase II trial. In: *Journal of neurology, neurosurgery, and psychiatry* 85 (11), S. 1198–1208. DOI: 10.1136/jnnp-2013-307282.

Oren, Ohad; Gersh, Bernard J.; Bhatt, Deepak L. (2020): Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. In: *The Lancet. Digital health* 2 (9), e486-e488. DOI: 10.1016/S2589-7500(20)30160-6.

Ouellette, Russell; Bergendal, Åsa; Shams, Sara; Martola, Juha; Mainero, Caterina; Kristoffersen Wiberg, Maria et al. (2018): Lesion accumulation is predictive of long-term cognitive decline in multiple sclerosis. In: *Multiple sclerosis and related disorders* 21, S. 110–116. DOI: 10.1016/j.msard.2018.03.002.

Park, Charlie C.; Brummer, Marijn E.; Sadigh, Gelareh; Saindane, Amit M.; Mullins, Mark E.; Allen, Jason W.; Hu, Ranliang (2022): Automated Registration and Color Labeling of Serial 3D Double Inversion Recovery MR Imaging for Detection of Lesion Progression in Multiple Sclerosis. In: *Journal of digital imaging*, S. 1–8. DOI: 10.1007/s10278-022-00737-1.

Peters, Sönke; Kellermann, Gesa; Watkinson, Joe; Gärtner, Friederike; Huhndorf, Monika; Stürner, Klarissa et al. (2024): AI supported detection of cerebral multiple sclerosis lesions decreases radiologic reporting times. In: *European journal of radiology* 178, S. 111638. DOI: 10.1016/j.ejrad.2024.111638.

Pham, N.; Hill, V.; Rauschecker, A.; Lui, Y.; Niogi, S.; Fillipi, C. G. et al. (2023): Critical Appraisal of Artificial Intelligence-Enabled Imaging Tools Using the Levels of Evidence System. In: *AJNR. American journal of neuroradiology* 44 (5), E21-E28. DOI: 10.3174/ajnr.A7850.

Polman, Chris H.; Reingold, Stephen C.; Banwell, Brenda; Clanet, Michel; Cohen, Jeffrey A.; Filippi, Massimo et al. (2011): Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. In: *Annals of neurology* 69 (2), S. 292–302. DOI: 10.1002/ana.22366.

Pucci, Eugenio; Giuliani, Giorgio; Solari, Alessandra; Simi, Silvana; Minozzi, Silvia; Di Pietrantonj, Carlo; Galea, Ian (2011): Natalizumab for relapsing remitting multiple sclerosis. In: *The Cochrane database of systematic reviews* (10), CD007621. DOI: 10.1002/14651858.CD007621.pub2.

Pugliatti, M.; Rosati, G.; Carton, H.; Riise, T.; Drulovic, J.; Vécsei, L.; Milanov, I. (2006): The epidemiology of multiple sclerosis in Europe. In: *European journal of neurology* 13 (7), S. 700–722. DOI: 10.1111/j.1468-1331.2006.01342.x.

Ramagopalan, Sreeram V.; Sadovnick, A. Dessa (2011): Epidemiology of multiple sclerosis. In: *Neurologic clinics* 29 (2), S. 207–217. DOI: 10.1016/j.ncl.2010.12.010.

Rathmann, Eiko; Hemkemeier, Pia; Raths, Susan; Grothe, Matthias; Mankertz, Fiona; Hosten, Norbert; Flessa, Steffen (2024): Changes in MRI Workflow of Multiple Sclerosis after Introduction of an AI-Software: A Qualitative Study. In: *Healthcare (Basel, Switzerland)* 12 (10). DOI: 10.3390/healthcare12100978.

Rice, G. P.; Incorvaia, B.; Munari, L.; Ebers, G.; Polman, C.; D'Amico, R.; Filippini, G. (2001): Interferon in relapsing-remitting multiple sclerosis. In: *The Cochrane database of systematic reviews* 2001 (4), CD002002. DOI: 10.1002/14651858.CD002002.

Riera, Rachel; Porfírio, Gustavo J. M.; Torloni, Maria R. (2016): Alemtuzumab for multiple sclerosis. In: *The Cochrane database of systematic reviews* 4 (4), CD011203. DOI: 10.1002/14651858.CD011203.pub2.

Rocca, Maria A.; Preziosa, Paolo; Barkhof, Frederik; Brownlee, Wallace; Calabrese, Massimiliano; Stefano, Nicola de et al. (2024): Current and future role of MRI in the diagnosis and prognosis of multiple sclerosis. In: *The Lancet regional health. Europe* 44, S. 100978. DOI: 10.1016/j.lanepe.2024.100978.

Rocca, Maria A.; Sormani, Maria Pia; Rovaris, Marco; Caputo, Domenico; Ghezzi, Angelo; Montanari, Enrico et al. (2017): Long-term disability progression in primary progressive multiple sclerosis: a 15-year study. In: *Brain : a journal of neurology* 140 (11), S. 2814–2819. DOI: 10.1093/brain/awx250.

Rovira, Àlex; Wattjes, Mike P.; Tintoré, Mar; Tur, Carmen; Yousry, Tarek A.; Sormani, Maria P. et al. (2015 Aug): Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis-clinical implementation in the diagnostic process. England.

Roy, Snehashis; Butman, John A.; Reich, Daniel S.; Calabresi, Peter A.; Pham, Dzung L. (2018): Multiple Sclerosis Lesion Segmentation from Brain MRI via Fully Convolutional Neural Networks.

Runmarker, B.; Andersen, O. (1993): Prognostic factors in a multiple sclerosis incidence cohort with twenty-five years of follow-up. In: *Brain : a journal of neurology* 116 (Pt 1), S. 117–134. DOI: 10.1093/brain/116.1.117.

Ruprecht, K.; Tumani, H. (2016): Liquordiagnostik bei Multipler Sklerose. In: *Der Nervenarzt* 87 (12), S. 1282–1287. DOI: 10.1007/s00115-016-0220-z.

Sadovnick, A. D.; Remick, R. A.; Allen, J.; Swartz, E.; Yee, I. M.; Eisen, K. et al. (1996): Depression and multiple sclerosis. In: *Neurology* 46 (3), S. 628–632. DOI: 10.1212/wnl.46.3.628.

Sahraian, M. A.; Radü, Ernst-Wilhelm (2008): MRI atlas of MS lesions. Limited ed. Berlin, Heidelberg: Springer.

Sawcer, Stephen; Franklin, Robin J. M.; Ban, Maria (2014): Multiple sclerosis genetics. In: *The Lancet. Neurology* 13 (7), S. 700–709. DOI: 10.1016/S1474-4422(14)70041-9.

Schmidt, Rudolf Manfred; Hoffmann, Frank; Faiss, Jürgen H.; Köhler, Wolfgang; Zettl, Uwe (2022): Multiple Sklerose. 8. Auflage. München, Deutschland: Elsevier.

Sieber, Victoria; Rusche, Thilo; Yang, Shan; Stieltjes, Bram; Fischer, Urs; Trebeschi, Stefano et al. (2024): Automated assessment of brain MRIs in multiple sclerosis patients significantly reduces reading time. In: *Neuroradiology* 66 (12), S. 2171–2176. DOI: 10.1007/s00234-024-03497-7.

Sima, Diana M.; Esposito, Giovanni; van Hecke, Wim; Ribbens, Annemie; Nagels, Guy; Smeets, Dirk (2021): Health Economic Impact of Software-Assisted Brain MRI on Therapeutic Decision-Making and Outcomes of Relapsing-Remitting Multiple Sclerosis Patients-A Microsimulation Study. In: *Brain sciences* 11 (12). DOI: 10.3390/brainsci11121570.

Simon, J. H.; Li, D.; Traboulsee, A.; Coyle, P. K.; Arnold, D. L.; Barkhof, F. et al. (2006): Standardized MR imaging protocol for multiple sclerosis: Consortium of MS Centers consensus guidelines. In: *AJNR. American journal of neuroradiology* 27 (2), S. 455–461.

Sormani, Maria Pia; Bruzzi, Paolo (2013): MRI lesions as a surrogate for relapses in multiple sclerosis: a meta-analysis of randomised trials. In: *The Lancet. Neurology* 12 (7), S. 669–676. DOI: 10.1016/S1474-4422(13)70103-0.

Sospedra, Mireia; Martin, Roland (2005): Immunology of multiple sclerosis. In: *Annual review of immunology* 23, S. 683–747. DOI: 10.1146/annurev.immunol.23.021704.115707.

Spagnolo, Federico; Depeursinge, Adrien; Schädelin, Sabine; Akbulut, Aysenur; Müller, Henning; Barakovic, Muhamed et al. (2023): How far MS lesion detection and segmentation are integrated into the clinical workflow? A systematic review. In: *NeuroImage. Clinical* 39, S. 103491. DOI: 10.1016/j.nicl.2023.103491.

Steinman, L. (1996): Multiple sclerosis: a coordinated immunological attack against myelin in the central nervous system. In: *Cell* 85 (3), S. 299–302. DOI: 10.1016/s0092-8674(00)81107-1.

Styner, Martin; Lee, Joohwi; Chin, Brian; Chin, Matthew S.; Commowick, Olivier; Tran, Hoai-Huong et al. (2008): 3D Segmentation in the Clinic: A Grand Challenge II: MS lesion segmentation. In: *The MIDAS Journal*. DOI: 10.54294/lmkqvm.

Sweeney, Melanie D.; Zhao, Zhen; Montagne, Axel; Nelson, Amy R.; Zlokovic, Berislav V. (2019): Blood-Brain Barrier: From Physiology to Disease and Back. In: *Physiological reviews* 99 (1), S. 21–78. DOI: 10.1152/physrev.00050.2017.

Tan, I. Leng; van Schijndel, Ronald A.; Fazekas, Franz; Filippi, Massimo; Freitag, Peter; Miller, David H. et al. (2002): Image registration and subtraction to detect active T(2) lesions in MS: an interobserver study. In: *Journal of neurology* 249 (6), S. 767–773. DOI: 10.1007/s00415-002-0712-6.

Thompson, Alan J.; Banwell, Brenda L.; Barkhof, Frederik; Carroll, William M.; Coetzee, Timothy; Comi, Giancarlo et al. (2018): Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. In: *The Lancet. Neurology* 17 (2), S. 162–173. DOI: 10.1016/S1474-4422(17)30470-2.

Tintore, Mar; Rovira, Àlex; Río, Jordi; Otero-Romero, Susana; Arrambide, Georgina; Tur, Carmen et al. (2015): Defining high, medium and low impact prognostic factors for developing multiple sclerosis. In: *Brain : a journal of neurology* 138 (Pt 7), S. 1863–1874. DOI: 10.1093/brain/awv105.

Traboulsee, Anthony (2004): MRI: role in optimising treatment. In: *Journal of neurology* 251 Suppl 5, v36-v41. DOI: 10.1007/s00415-004-1506-9.

Tran, Philippe; Thoprakarn, Urielle; Gourieux, Emmanuelle; Dos Santos, Clarisse Longo; Cavedo, Enrica; Guizard, Nicolas et al. (2022): Automatic segmentation of white matter hyperintensities: validation and comparison with state-of-the-art methods on both Multiple Sclerosis and elderly subjects. In: *NeuroImage. Clinical* 33. DOI: 10.1016/j.nicl.2022.102940.

Trojano, M.; Paolicelli, D.; Bellacosa, A.; Cataldo, S. (2003): The transition from relapsing-remitting MS to irreversible disability: clinical evaluation. In: *Neurological sciences : official journal of the Italian Neurological Society and of*

the Italian Society of Clinical Neurophysiology 24 Suppl 5, S268-70. DOI: 10.1007/s10072-003-0171-6.

Tsagkas, Charidimos; Magon, Stefano; Gaetano, Laura; Pezold, Simon; Naegelin, Yvonne; Amann, Michael et al. (2019): Preferential spinal cord volume loss in primary progressive multiple sclerosis. In: *Multiple sclerosis (Houndmills, Basingstoke, England)* 25 (7), S. 947–957. DOI: 10.1177/1352458518775006.

Uher, Tomas; Vaneckova, Manuela; Sobisek, Lukas; Tyblova, Michaela; Seidl, Zdenek; Krasensky, Jan et al. (2017): Combining clinical and magnetic resonance imaging markers enhances prediction of 12-year disability in multiple sclerosis. In: *Multiple sclerosis (Houndmills, Basingstoke, England)* 23 (1), S. 51–61. DOI: 10.1177/1352458516642314.

Valverde, Sergi; Cabezas, Mariano; Roura, Eloy; González-Villà, Sandra; Pareto, Deborah; Vilanova, Joan C. et al. (2017): Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. In: *NeuroImage* 155, S. 159–168. DOI: 10.1016/j.neuroimage.2017.04.034.

Valverde, Sergi; Salem, Mostafa; Cabezas, Mariano; Pareto, Deborah; Vilanova, Joan C.; Ramió-Torrentà, Lluís et al. (2019): One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. In: *NeuroImage. Clinical* 21, S. 101638. DOI: 10.1016/j.nicl.2018.101638.

van Hecke, Wim; Costers, Lars; Descamps, Annabel; Ribbens, Annemie; Nagels, Guy; Smeets, Dirk; Sima, Diana M. (2021): A Novel Digital Care Management Platform to Monitor Clinical and Subclinical Disease Activity in Multiple Sclerosis. In: *Brain sciences* 11 (9). DOI: 10.3390/brainsci11091171.

Waller, Joseph; O'Connor, Aisling; Rafaat, Eleeza; Amireh, Ahmad; Dempsey, John; Martin, Clarissa; Umair, Muhammad (2022): Applications and challenges of artificial intelligence in diagnostic and interventional radiology. In: *Polish journal of radiology* 87, e113-e117. DOI: 10.5114/pjr.2022.113531.

Walton, Clare; King, Rachel; Rechtman, Lindsay; Kaye, Wendy; Leray, Emmanuelle; Marrie, Ruth Ann et al. (2020): Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS, third edition. In: *Multiple sclerosis (Houndmills, Basingstoke, England)* 26 (14), S. 1816–1821. DOI: 10.1177/1352458520970841.

Wang, Shui-Hua; Tang, Chaosheng; Sun, Junding; Yang, Jingyuan; Huang, Chenxi; Phillips, Preetha; Zhang, Yu-Dong (2018): Multiple Sclerosis Identification by 14-Layer Convolutional Neural Network With Batch Normalization, Dropout, and Stochastic Pooling. In: *Frontiers in neuroscience* 12, S. 818. DOI: 10.3389/fnins.2018.00818.

Wang, W.; van Heerden, J.; Tacey, M. A.; Gaillard, F. (2017): Neuroradiologists Compared with Non-Neuroradiologists in the Detection of New Multiple Sclerosis Plaques. In: *AJNR. American journal of neuroradiology* 38 (7), S. 1323–1327. DOI: 10.3174/ajnr.A5185.

Wattjes, M. P.; Lutterbey, G. G.; Gieseke, J.; Träber, F.; Klotz, L.; Schmidt, S.; Schild, H. H. (2007): Double inversion recovery brain imaging at 3T: diagnostic value in the detection of multiple sclerosis lesions. In: *AJNR. American journal of neuroradiology* 28 (1), S. 54–59.

Wattjes, Mike P.; Rovira, Àlex; Miller, David; Yousry, Tarek A.; Sormani, Maria P.; Stefano, Maria P. de et al. (2015 Oct): Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis--establishing disease prognosis and monitoring patients. England.

Weiner, Howard L. (2009): The challenge of multiple sclerosis: how do we cure a chronic heterogeneous disease? In: *Annals of neurology* 65 (3), S. 239–248. DOI: 10.1002/ana.21640.

Weinshenker, B. G.; Bass, B.; Rice, G. P.; Noseworthy, J.; Carriere, W.; Baskerville, J.; Ebers, G. C. (1989): The natural history of multiple sclerosis: a geographically based study. I. Clinical course and disability. In: *Brain : a journal of neurology* 112 (Pt 1), S. 133–146. DOI: 10.1093/brain/112.1.133.

Weinshenker, B. G.; O'Brien, P. C.; Petterson, T. M.; Noseworthy, J. H.; Lucchinetti, C. F.; Dodick, D. W. et al. (1999): A randomized trial of plasma exchange in acute central nervous system inflammatory demyelinating disease. In: *Annals of neurology* 46 (6), S. 878–886. DOI: 10.1002/1531-8249(199912)46:6<878::aid-ana10>3.0.co;2-q.

Wildemann, Brigitte; Oschmann, Patrick; Reiber, Hansotto; Brettschneider, J.; Ebinger, F.; Gasser, T. et al. (Hg.) (2006): Neurologische Labordiagnostik. 1. Auflage. Stuttgart: Georg Thieme Verlag (Referenz-Reihe Neurologie: Methoden). Online verfügbar unter http://ebooks.thieme.de/9783131365910/1.

Winder, Anthony J.; Stanley, Emma am; Fiehler, Jens; Forkert, Nils D. (2024): Challenges and Potential of Artificial Intelligence in Neuroradiology. In: *Clinical neuroradiology* 34 (2), S. 293–305. DOI: 10.1007/s00062-024-01382-7.

Wolff, Justus; Pauling, Josch; Keck, Andreas; Baumbach, Jan (2020): The Economic Impact of Artificial Intelligence in Health Care: Systematic Review. In: *Journal of medical Internet research* 22 (2), e16866. DOI: 10.2196/16866.

Wray, Sibyl; Then Bergh, Florian; Wundes, Annette; Arnold, Douglas L.; Drulovic, Jelena; Jasinska, Elzbieta et al. (2022): Efficacy and Safety Outcomes with Diroximel Fumarate After Switching from Prior Therapies or Continuing on DRF: Results from the Phase 3 EVOLVE-MS-1 Study. In: *Advances in therapy* 39 (4), S. 1810–1831. DOI: 10.1007/s12325-022-02068-7.

Yamamoto, Daisuke; Arimura, Hidetaka; Kakeda, Shingo; Magome, Taiki; Yamashita, Yasuo; Toyofuku, Fukai et al. (2010): Computer-aided detection of multiple sclerosis lesions in brain magnetic resonance images: False positive reduction scheme consisted of rule-based, level set method, and support vector machine. In: *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society* 34 (5), S. 404–413. DOI: 10.1016/j.compmedimag.2010.02.001.

Yan, X-B; Zhao, Y-F; Yang, Y-M; Wang, N.; He, B-Z; Qiu, X-T (2019): Impact of astrocyte and lymphocyte interactions on the blood-brain barrier in multiple sclerosis. In: *Revue neurologique* 175 (6), S. 396–402. DOI: 10.1016/j.neurol.2018.12.006.

Zaharchuk, G.; Gong, E.; Wintermark, M.; Rubin, D.; Langlotz, C. P. (2018): Deep Learning in Neuroradiology. In: *AJNR. American journal of neuroradiology* 39 (10), S. 1776–1784. DOI: 10.3174/ajnr.A5543.

Zhang, Jian; Shi, Shengliang; Zhang, Yueling; Luo, Jiefeng; Xiao, Yousheng; Meng, Lian; Yang, Xiaobo (2017): Alemtuzumab versus interferon beta 1a for relapsing-remitting multiple sclerosis. In: *The Cochrane database of systematic reviews* 11 (11), CD010968. DOI: 10.1002/14651858.CD010968.pub2.

Zhao, Yue; Guo, Shuxu; Luo, Min; Shi, Xue; Bilello, Michel; Zhang, Shaoxiang; Li, Chunming (2018): A level set method for multiple sclerosis lesion segmentation. In: *Magnetic resonance imaging* 49, S. 94–100. DOI: 10.1016/j.mri.2017.03.002.

Zijdenbos, Alex P.; Forghani, Reza; Evans, Alan C. (2002): Automatic "pipeline" analysis of 3-D MRI data for clinical trials: application to multiple sclerosis. In: *IEEE transactions on medical imaging* 21 (10), S. 1280–1291. DOI: 10.1109/TMI.2002.806283.

Zivadinov, Robert; Uher, Tomas; Hagemeier, Jesper; Vaneckova, Manuela; Ramasamy, Deepa P.; Tyblova, Michaela et al. (2016): A serial 10-year follow-up study of brain atrophy and disability progression in RRMS patients. In: *Multiple sclerosis (Houndmills, Basingstoke, England)* 22 (13), S. 1709–1718. DOI: 10.1177/1352458516629769.

9 Erklärung zum Eigenanteil

Die Arbeit wurde in der Abteilung für Diagnostische und Interventionelle Neuroradiologie im Department Radiologie des Universitätsklinikums Tübingen unter Betreuung von zunächst Prof. Thomas Nägele und dann Prof. Ulrike Ernemann durchgeführt.

Die Konzeption der Studie erfolgte durch Prof. Benjamin Bender und Dr. Tobias Lindig.

Prof. Bender hat mich bei der Methoden- und Toolwahl beraten, die ich anschließend selbstständig angewandt habe.

Die visuelle Bewertung wurde selbstständig durch mich, Prof. Bender und Dr. Oliver Krüger durchgeführt. Dabei kam in einer der Auswertungsrunden ergänzend ein KI-basiertes Tool zur automatisierten Läsionsdetektion zum Einsatz.

Die Versuche wurden von mir in Zusammenarbeit mit Prof. Dr. Benjamin Bender und Dr. Oliver Krüger durchgeführt.

Die statistische Auswertung erfolgte nach Beratung durch das Institut für Biometrie selbstständig durch mich.

Literaturrecherche, Visualisierung und Erstellung des finalen Manuskripts habe ich selbstständig durchgeführt und ich versichere, keine weiteren als die von mir angegebenen Quellen verwendet zu haben.

Tübingen, den 07.09.2025

10 Veröffentlichungen

Finn S, Krüger O, Lindig T, Ernemann U, Bender B. Einfluss einer automatisierten Läsionssegmentierung in MRT-Verlaufsuntersuchungen auf die Erkennungsrate von progredienten MS Fällen. Vortrag im Rahmen der Hauptsitzung 6: Neurodegeneration & Immunologie, 59. Jahrestagung der Deutschen Gesellschaft für Neuroradiologie e.V.

11 Danksagung

Mein besonderer Dank gilt Prof. Dr. Bender, der mich während meiner Promotionszeit mit großem Engagement, fachlicher Kompetenz und Unterstützung begleitet hat.

In tiefer Dankbarkeit erinnere ich mich an Prof. Dr. Nägele, der mir als Doktorvater mit großem Interesse und wissenschaftlicher Expertise zur Seite stand und mich zu dieser Arbeit ermutigte. Sein plötzlicher Tod war ein großer Verlust. Nach seinem Tod übernahm Prof. Dr. Ernemann die Doktormutterschaft, wofür ich sehr dankbar bin, und führte das Promotionsverfahren weiter.

Mein herzlicher Dank gilt außerdem meiner Familie und meinen Freunden, die mich in dieser Zeit durch ihre Geduld, Ermutigung und Unterstützung getragen haben.